DOCUMENT RESUME

ED 206 717 TM 810 623

AUTHOR Nicewander, W. Alan; And Others
TITLE A Comparison of Two Methods of Interactive Testing. Final Report.
INSTITUTION Oklahoma Univ., Norman. Dept. of Psychology.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [80]
GRANT NIE-G-79-0145
NOTE 79p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Computer Assisted Testing; Guessing (Tests); Item Banks; Latent Trait Theory; Mathematical Models; *Multiple Choice Tests; Questioning Techniques; *Test Format
IDENTIFIERS *Respond Until Correct Method; *Tailored Testing

ABSTRACT

Two methods of interactive, computer-assisted testing methods for multiple-choice items were compared with each other and with conventional multiple-choice tests. The interactive testing methods compared were tailored testing and the respond-until-correct (RUC) item response method. In tailored testing, examinee ability is successively estimated during the course of test administration, and optimal items, based on the most current ability estimate, are administered to maximize the test's information for the examinee's ability level. The RUC response mode requires examinees to respond repeatedly to mulitple-choice items until the correct answer is discovered. RUC tests consist of a fixed set of items for all examinees. The comparisons among testing methods--tailored, RUC and conventional--were analytical, rather than empirical, and the primary index for the comparisons was the average test information. The analytic comparisons of the testing methods indicated that in certain situations, RUC tests can provide greater average information than tailored tests, and furthermore, there are circumstances in which conventional tests provide average information that is essentially equal to that of the tailored tests. However, in terms of providing adequate information across nearly all ability levels, tailored tests have no competitor. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED2067717

A Comparison of Two Methods of Interactive Testing

W. Alan Nicewander, Horng-Shing Chang
and
Evelyn N. Doody

Department of Psychology
University of Oklahoma

Final Report
NIE Grant NIE-G-79-0145
National Institute of Education
Washington, D.C. 20208

TM 810 623

Abstract

Two methods of interactive, computer-assisted testing methods for multiple-choice items were compared with each other and with conventional multiple-choice tests. The interactive testing methods compared were tailored testing and the respond-until-correct (RUC) item response method. In tailored testing, examinee ability is successively estimated during the course of test administration, and optimal items, based on the most current ability estimate, are administered. In tailored testing examinees are given different tests which are selected in a manner to maximize the test's information for the examinee's ability level. The RUC response mode requires examinees to respond repeatedly to multiple-choice items until the correct answer is discovered. RUC tests consist of a fixed set of items for all examinees.

The comparisons among testing methods -- tailored, RUC and conventional -- were analytical, rather than empirical, and the primary index for the comparisons was the average test information. The analytic comparisons of the testing methods indicated that in certain situations, RUC tests can provide greater average information than tailored tests, and furthermore, there are circumstances in which conventional tests provide average information that is essentially equal to that of the tailored tests. However, in terms of providing adequate information across nearly all ability levels, tailored tests have no competitor.

Acknowledgements

I wish to thank the National Institute of Education for their support of this basic research project. Funds for basic research are difficult to obtain, and NIE's decision to aid our research efforts are deeply appreciated. I hope that we have produced something of value to other researchers in the areas of applied and theoretical psychometrics.

I express my gratitude to Horng-Shing Chang and Evelyn N. Doody whose skill and dedication made this project possible. A special note of thanks is owed to Dr. Albert Schwarzkoph of the Department of Mathematics at Oklahoma University and Dr. Mark Reckase of the University of Missouri for their extremely valuable advice and wise counsel.

I am bound to record my gratitude to Tom Harrison and Nel Brinlee, Office of Grants and Contracts, and Dr. N. Jack Kanak, Chairman, Department of Psychology, for their special help and support.

A Comparison of Two Interactive Testing Methods

The increasing availability and affordability of computer terminals and mini-micro computer systems has made possible the administration of tests in an interactive mode. Computer administered tests have at least two advantages over paper-and-pencil tests: 1) Immediate feedback concerning performance allows the test to serve as a teaching device; and 2) the amount of information about examinee ability can be increased relative to the information obtained from paper-and-pencil tests. We propose to examine two methods of interactive testing--tailored testing and respond-until-correct (RUC) testing. Of these two methods, tailored testing is the newest and probably the best known (Urry, 1974; Samejima, 1977). RUC testing dates back at least as far as 1926 with the work of Pressey (1926). More recently Kane and Moloney (1978) and Nicewander, Chang, Doody and Schwarzkopf (1980) have presented theoretical evidence demonstrating the superiority of the RUC testing-scoring method over the usual binary scoring method.

Tailored testing is an outgrowth of latent ability theories of mental test scores (Lord, 1952; Lord & Novick, 1968 and Birnbaum, 1968). In latent ability test theory it is assumed that examinees are described by a one-dimensional ability random variable, θ, scaled to have mean zero and variance one. Further it is assumed that the probability that an item is answered correctly is a function of the examinee's ability, θ, and the difficulty, discrimination and guessing parameters for the item. If we let U_g be the binary item score for the item g, then $P_g(\theta)$, the condition probability of answering the item correctly, is given by:

$$(1) \qquad P_g(\theta) = p(U_g = 1|\theta) = c_g + (1-c_g)F[L_g(\theta)],$$

where $L_g(\theta) = a_g(\theta - b_g)$; c_g is the guessing parameter (probability), and a_g and b_g are, respectively, the item discrimination and difficulty parameters. In (1) F represents either the normal cummulative density function (CDF) or the logistic CDF. Since there is no psychological reason to prefer either the normal or logistic CDF, we chose to work with a normal CDF because the numerical integration required was easier (for many problems the converse is true). Therefore, the three-parameter normal ogive model to be used in this investigation can be written as:

$$P_g(\theta) = c_g + (1-c_g)\,\phi[L_g(\theta)], \tag{2}$$

where $\phi[L_g(\theta)]$ is defined as,

$$\phi[L_g(\theta)] = \int_{-\infty}^{L_g(\theta)} \phi(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{L_g(\theta)} \exp(-t^2/2)dt. \tag{3}$$

In this model for test scores the information function (See Birnbaum, 1968, Ch. 20) for an item score is given by

$$I(\theta, U_g) = \frac{\partial}{\partial\theta}[P_g(\theta)]^2 \,/\, \sigma^2(U_g|\theta) = \{(1-c_g)\, a_g\, \phi[L_g(\theta)]\}^2 \,/\, P_g(\theta)\, Q_g(\theta) \tag{4}$$

where $I(\theta, U_g)$ is the information value of item g for persons of ability = θ; $\phi(t)$ is the normal probability density at t, and $Q_g(\theta) = 1-P_g(\theta)$. The information function ranges between zero and .6365 a_g^2 (When $c_g = 0$). In practice it is rare to find items with a_g exceeding 2; therefore, the practical limits of the information function are zero and 4(.6365) = 2.5. The information function for an item is an extremely useful measure of the operating characteristics of an item, indicating, conditional on ability, how accurately ability can be

estimated from the item score (Samejima, 1977). Additionally, the information function for the total test is simply the sum of the individual item information functions. Other measures of a test's characteristics such as the linear correlation between observed test score and ability (the test validity), and the square of the non-linear correlation coefficient (eta^2) between test score and ability (the test reliability), are not easily found by aggregating the corresponding item-ability correlation coefficients, and require that one specifies a probability distribution for θ.

In a nutshell, tailored testing utilizes latent ability theory to choose optimal items for each examinee in the following manner:

i) an "item-bank" must be available containing items with known (estimated) values of a_g, b_g and c_g.

ii) the examinee is administered an item, and his/her ability is estimated via a maximum likelihood fitting of the latent ability model.

iii) the information values are computed for all remaining items at the estimated ability level.

iv) the item with the highest information (at the estimated ability level) is administered to the subject.

v) repeat (iii) and (iv) until a predetermined test information value is achieved or a fixed number of items has been administered.

vi) the tailored test score is the final maximum likelihood estimate of the examinee's ability, $\hat{\theta}$.

This brief description of tailored testing is oversimplified in that maximum likelihood estimates of examinee ability are not possible mathematically until the examinee has answered at least one item correctly and has missed at least one item; therefore, until this state of affairs is attained, item selection is accomplished by one of several sub-optimal item selection procedures.

The advantage of tailored-tests over the standard type of test with identical items for all examinees (called "fixed tests" here) can be considerable

in terms of information, especially if the available item bank is extensive in terms of item difficulty and contains items with a_g values in the range .8 and higher (Urry, 1977). Since tailored tests must be computer administered, the correct answer to every item can be revealed during the testing session, thus making a tailored test the teacher's as well as the psychometrician's delight.

Tailored testing also gives rise to some rather difficult problems. Every examinee takes a different test; therefore, serious legal problems could arise. The psychometric superiority of tailored testing is not intuitively obvious to the layman. Furthermore, the number-correct test score is useless for tailored tests, and one will certainly be confronted by examinees having identical number-correct scores, but drastically different estimates of θ due to differences in the difficulty of the tailored tests. The latent-ability model generating the ability estimates is difficult to explain to professionals, and will appear totally recondite in the eyes of examinees.

Nicewander, Chang, Doody and Schwarzkopf (1980) have proposed a model for RUC item scores that is a modification of the three-parameter normal ogive model given in Equation (2). Before presenting this model some additional background is necessary. First rewrite (2) as

$$(5) \qquad P_g(\theta) = \Phi[L_g(\theta)] + c_g\,\{1-\Phi[L_g(\theta)]\}, = \Phi(\chi_g) + c_g\,[1-\Phi(\chi_g)],$$

where $\chi_g \equiv L_g(\theta)$. This model has been justified in the past in two quite distinct manners: 1) As a mathematical formulation of the Knowledge or Random Guessing (KRG) principle which states that the examinee "knows" the answer to the g^{th} item, with probability $\Phi(\chi_g)$, or doesn't know the answer (with probability equal to $1 - \Phi(\chi_g)$) but guesses randomly and is correct with probability

equal $c_g = 1/A$, and 2) More recently, investigators have used (5) as a falsifiable model which may be fitted to data by estimating a_g, b_g and c_g, and which may or may not describe the observed item responses. In this interpretation of the model, c_g is treated as the lower asymptote for the ogival function given by (5), and is simply another parameter to be estimated from the observed data. In order to derive a latent ability model for RUC item responses, it will be convenient to give another type of cognitive (as opposed to empirical) interpretation of the equation for $P_g(\theta)$ given in (5). First we reject the KRG principle as being untenable. [For example: Prater and Nicewander (1980) found that examinees, when given second attempts at incorrectly answered five-choice vocabulary items, were correct with probability .39; whereas, KRG predicts that the second-attempt probability should be .25.] Secondly, let $\Phi(\chi_g) = \Phi[L_g(\theta)]$ be the probability that an examinee with ability = θ selects a strategy that leads to the correct answer; therefore, $1 - \Phi(\chi_g)$ is the probability that the examinee selects a strategy that does not lead to the correct response. The parameter c_g is interpreted as a guessing probability which is, in general, not equal to $1/A$. We may summarize our interpretation of Equation (5) as follows. An examinee can obtain the correct answer to an item by 1) selecting a strategy that leads to the correct answer, or 2) by selecting a strategy that does not lead to the correct answer and guessing correctly. For example, when confronted by a vocabulary item requesting the synonym for a particular word, the examinee might consider strategies such as, "search semantic memory and determine if a synonym can be found that matches one of the alternatives", or "educe the Latin stem for the word, and check the alternatives to see if one of them is consonant with the educed Latin stem", or (for extremely difficult, low-frequency usage words) "check the alternative ensemble; if there is a recognizable word, choose it

as the correct answer", and so forth. For problems involving mathematics it is fairly easy to compile a lengthy list of strategies that an examinee might consider adopting in order to solve a particular problem -- some will lead to the correct response; others will not. The greater the examinee's ability, the higher the probability of choosing a strategy leading to the correct answer.

Define r as the number of attempts made in discovering the correct answer to an A-choice item administered under the RUC format. Let $Y_g=r$ be the RUC score for the g^{th} item, and $P_{gr}(\theta) \equiv p(Y_g=r|\theta)$ represent the conditional probability that r attempts are necessary for answering the g^{th} item by persons with ability equal to θ. Our model proposes that the conditional probability that one attempt is required for answering a particular item is given by

$$(6) \qquad p(Y_g=1|\theta) = P_{g1}(\theta) = \Phi(x_g) + c_{g1}[1-\Phi(x_g)] = c_{g1} + (1-c_{g1})\Phi(x_g)$$

which is identical to the probability that the item is correct under "one-zero" (binary) scoring (with the exception that c_g is further subscripted to reflect the number of attempts).

We further propose that it is reasonable to assume that the conditional probability that two attempts are required to answer the g^{th} item is approximated by

$$(7) \qquad p(Y_g=2|\theta)=P_{g2}(\theta)= \Phi(x_g)[1-\Phi(x_g)](1-c_{g1})+[1-\Phi(x_g)]^2(1-c_{g1})c_{g2}$$

where c_{g2} is the guessing probability for the second attempt. Equation (7) is now justified in terms of the behavior being modeled. Recall that we are assuming that $\Phi(x_g)$ is the probability that an examinee with ability equal to θ selects a strategy that leads to the correct answer. It is convienient (and probably not a serious distortion of reality) to assume that the number of

possible strategies is sufficiently large so that $\phi(x_g)$ remains constant from one attempt to another. That is to say, the probability of selecting an efficacious strategy on the r^{th} attempt at answering an item is the same under sampling of strategies with-or-without replacement. Equation (7) can then be interpreted as follows: An examinee can be correct on the second attempt by; 1) Choosing an incorrect strategy and quessing incorrectly on the first attempt, but choosing a successful strategy on the second attempt -- this state of affairs has probability equal to $\phi(x_g)[1-\phi(x_g)](1-c_{g1})$, or 2) Choosing an incorrect strategy on <u>both</u> attempts, guessing incorrectly on the first attempt and guessing correctly on the second attempt -- this sequence quite clearly has probability equal to $[1-\phi(x_g)]^2(1-c_{g1})c_{g2}$.

This logic is carried one more step before presenting a general expression for the conditional probability distribution for the RUC score, Y_g. For the conditional probability of obtaining the correct answer on the third attempt at answering the g^{th} item, we assume that

$$(8) \quad p(Y=3|\theta) = P_{g3}(\theta) = \phi(X_g)[1-\phi(X_g)]^2(1-c_{g1})(1-c_{g2})+[1-\phi(X_g)]^3(1-c_{g1})(1-c_{g2})c_{g3},$$

where c_{g3} is the guessing probability for the third attempt. Briefly, (8) implies that an examinee can be correct on the third attempt at an item by: 1) choosing incorrect strategies and quessing incorrectly on attempts one and two, but choosing an efficacious strategy on the third attempt, or 2) Choosing incorrect strategies on attempts one, two and three, guessing incorrectly on the first two attempts, but guessing correctly on the third.

It is now possible to write a general expression for the conditional probability distribution of the RUC score, Y_g, for an A-choice item as

$$(9) \quad p(Y_g=r|\theta)=P_{gr}(\theta)=[1-\phi(X_g)]^{r-1}\prod_{i=1}^{r}(1-c_{g,i-1})\{\phi(X_g)+c_{gr}[1-\phi(X_g)]\}$$

where $c_{g0} \equiv 0$, and $r=1, 2, \ldots, A$.

The model given in (9) for RUC scores is only one of several approaches that could be taken in modeling examinee behavior in the situation where multiple-choice items are responded to repeatedly until the correct answer is discovered. We believe that our proposed model represents one of the simplest possible (in terms of the number of parameters), and is a rather straightforward extension of a three-parameter normal ogive model for binary scoring. In (9) notice that, for an item with A-choices, there are A-1+2 = A+1 free parameters. For example, a five-choice item administered under the proposed RUC model has a_g, b_g, c_{g1}, c_{g2}, c_{g3} and c_{g4} as free parameters (c_{g5}=1, always).

Another approach that one could adopt in modeling the RUC item score, Y_g, is to assume that the examinee either knows the correct answer with certainty (and is correct with probability equal to one) or sorts the alternatives into two subsets -- "plausible" and "implausible" -- and chooses alternatives (possibly at random) from the "plausible" subset. We are currently working on models of this genre; however, it is very difficult to derive one that is mathematically tractable. Other promising methods for arriving at a theory for RUC item scores are suggested by the recent work of Samejima (1979) wherein she derives a family of models for multiple-choice items which consider the item foils as well as the correct responses.

In the present research we limited ourselves to the study of five-choice items. In order to facilitate the derivation of the information function for the RUC item response model given in (9), it will be convenient to define a (linear) scoring function, $M_g(r)$, such that

(10) $$M_g(r) = 5-r,$$

where r is the number of attempts made in answering the g^{th} item. Using this scoring function, in conjunction with five-choice RUC items, the possible

item scores are 4, 3, 2, 1 and 0. The information function for the RUC item response model for five-choice items scored under $M_g(r) = 5-r$ is given by

$$(11) \quad I(\theta,Y_g)=[\frac{\partial}{\partial\theta}\mu(Y_g|\theta)]^2/\sigma^2(Y_g|\theta) = [K_g a_g \phi(\chi_g)]^2/\sigma^2(Y_g|\theta),$$

where $\mu(Y_g|\theta)$ and $\sigma^2(Y_g|\theta)$ are, respectively, the conditional mean and variance of Y_g. Furthermore,

$$(12) \quad K_g = -[4\prod_{r=1}^{4}(1-c_{gr})\phi^3(-\chi_g)+3\prod_{r=1}^{3}(1-c_{gr})\phi^2(-\chi_g)+2\prod_{r=1}^{2}(1-c_{gr})\phi(-\chi_g)+1-c_{g1}],$$

where $\phi^n(-\chi_g) \equiv [\phi(-\chi_g)]^n = [1-\phi(\chi_g)]^n$.

The ratio of the information functions for RUC items given in (11) to that of binary items is

$$(13) \quad \frac{K_g^2}{(1-c_g)^2} \cdot \frac{\sigma^2(U_g|\theta)}{\sigma^2(Y_g|\theta)},$$

and one will note that $I(Y_g,\theta)>I(U_g|\theta)$ if $K_g^2>(1-c_g)^2\sigma^2(Y_g|\theta)/\sigma^2(U_g|\theta)$. However, cognizance of this inequality is of little value in determining under what conditions the RUC score, Y_g, will yield greater information than U_g, the binary score. It is clear that the pattern of guessing probabilities across attempts is the crucial determinant as to whether or not the RUC scoring system will show superiority in information relative to binary scoring. For example if only c_{g1} and c_{g2} are non-zero, then the information in the RUC item score must be equal to that of the binary score -- in this instance Y_g is simply a linear transform of U_g, the binary score. For five different patterns of the c_{gr} (to be described later) Nicewander, Chang, Doody and Schwarzkopf (1980) found, for all choices of a_g studied, that the information in the RUC item score was always at least as great as the binary score; for certain combinations of a_g and θ, the information in the RUC score was 150 percent greater than that of the binary score.

Guessing Patterns Utilized in the Investigation

The difference in information between items scored one-or-zero and items scored under RUC depends principally on the pattern of the guessing probabilities across repeated attempts at answering a multiple-choice item. The variety of these "guessing patterns" is, of course, infinite; consequently, we chose to work with five representative guessing patterns for RUC items. These five guessing patterns were selected so as to represent typical situations apt to exist in the alternative ensemble for a multiple-choice item. Table 1 lists the five guessing patterns -- called G_1, G_2, G_3, G_4 and G_5 -- that were used in Phases I and II of this inquiry. The pattern G_1 is the random guessing pattern; that is to say, the values for the c_{gr} are equal to the reciprocals of the number of remaining alternatives on attempt r in a RUC format. G_2 gives values of the guessing probabilities that are considerably higher than those expected under random guessing. The pattern G_3 was designed to simulate the case where low ability subjects are "misinformed". G_3 generates conditional probability distributions for Y_g in which examinees of low ability have a relatively high probability of choosing an incorrect alternative on the first attempt. Pattern G_4 was included to illustrate the effects of an item with a non-functioning alternative, and finally G_5 is a pattern of guessing for an item having two non-functioning alternatives, where a non-functioning alternative is defined as an alternative never selected by examinees as the correct answer. Another way of describing G_4 and G_5 is to indicate that under G_4 an item is always answered in four or fewer attempts, and under G_5 the correct alternative is always discovered in three or fewer attempts. The first element of the guessing patterns, c_{gr}, was used as the guessing parameter for the binary items comprising either tailored or conventional tests.

Table 1

Five Guessing Patterns for
RUC Model Applied to 5-choice Items

Pattern	Guessing Probability on r^{th} Attempt				
	c_{g1}	c_{g2}	c_{g3}	c_{g4}	c_{g5}
G_1	.20	.25	.33	.50	1.00
G_2	.40	.50	.67	.99	1.00
G_3	.10	.25	.33	.50	1.00
G_4	.24	.32	.49	1.00	--
G_5	.50	.80	1.00	--	--

Phase I. A Comparison of Perfect Tailored Tests, RUC Tests and Conventional Tests Using Items From Ideal Item Banks.

This section of the investigation consisted of a series of idealized experiments. These experiments may be viewed as a comparison of the three types of testing under limiting conditions. In order to accomplish these idealized experiments, we defined three concepts:

1) Ideal Item Bank. An infinitely large set of items with fixed values of a_g and guessing pattern (hence, fixed c_g for binary items), and differing only in difficulty (b_g). In an ideal item bank an item with any value of b_g is available in unlimited quantity.

2) Perfect Tailored Tests. A perfect tailored test is that set of items that maximimizes the test information at a known ability. Perfect tailored tests may be contrasted with imperfect tailored tests in which the ability of examinees is unknown and must be successively approximated.

3) Types I, II and III RUC and Conventional Tests. These tests are composed of a fixed set of items (binary or RUC) designed to be optimal for certain portions of the ability distribution. Type I tests, composed of either a fixed set of binary or RUC items, are defined as those item collections which maximize the average test information for the complete ability distribution. Type II RUC or binary tests are sets of items which maximize the conditional average test information function for examinees whose ability is average or above. Type III RUC or binary tests consist of a set of fixed items selected to maximimize the conditional average test information for abilities average and below.

For the comparison of the three methods of testing (Tailored, RUC and conventional), 25 ideal item banks were defined by taking all combinations of five guessing patterns and five values of a_g(.6, .8, 1.0, 1.5 and 2.0). In other words, the combination of guessing pattern G_1 and a_g = .6 defines an ideal item bank from which perfect tailored tests and Type I, II and III RUC and conventional (binary) tests may be constructed. Note that only the first element of a guessing pattern, c_{g1}, is relevant for the binary items used in the perfect-tailored and conventional tests.

The test information function for a perfect tailored test will be constant across all ability levels. This is true because, for any ability, the maximum value of the information function for a binary item is determined by a_g and $c_g = c_{g1}$; whether or not the maximum is attained, depends upon b_g. Since any value of b_g is available in an ideal item bank, the maximum information is attainable for all θ. In order to find the maximum (constant) information value for the perfect tailored tests drawn from a specific ideal item bank, the maximum information was found at $\theta = 0$ using a simple computer-search algorithm that manipulated b_g until the maximum value of $I(\theta, U_g)$ was found. The maximum value of $I(\theta, U_g)$ was multiplied by 10 in order to provide a test information function for a 10-item tailored test -- this value served as the constant test information function for the perfect tailored tests of 10 items each. This procedure was repeated for all ideal item banks in order to derive the test information for perfect tailored tests.

The Type I RUC and conventional tests drawn from a specific ideal item bank were constructed by finding the value of b_g (the only free item parameter in an ideal bank) which maximized,

$$(14) \qquad \mu[I(\theta,S_g)] = \int_{-\infty}^{+\infty} I(\theta,S_g)\phi(\theta)d\theta,$$

the (unconditional) average information. S_g represents either U_g, the binary item score, or Y_g, the RUC item score. The integration in (14) was done numerically using Gaussian-Hermite quadrature. A computer-search procedure was used to find the values of b_g that maximized (14) for both binary and RUC items. The items with b_g values that maximized $\mu[I(\theta, S_g)]$ for the RUC and binary scoring methods were "repeated" 10 times to produce fixed, Type I RUC and conventional tests equal in length to the perfect tailored tests. This process was repeated for each of the 25 ideal item banks. The test information functions were plotted and their averages (across ability) computed.

The construction of the Type II and III RUC and conventional tests was identical to that of the Type I tests with the exception that items for these tests were chosen so as to maximize the following conditional average information functions:

$$(15) \qquad \mu[I(\theta,S_g|\theta\geq 0)]=2\int_0^{+\infty} I(\theta,S_g)\phi(\theta)d\theta \ ,$$

for Type II RUC and binary items, and

$$(16) \qquad \mu[I(\theta,S_g|\theta\leq 0)]=2\int_{-\infty}^{0} I(\theta,S_g)\phi(\theta)d\theta \ ,$$

for Type III RUC and binary items. The integration in (15) and (16) was done via a modified Gaussian-Hermite quadrature procedure in which only the positive roots of the Hermite polynomials were used for integrating (15), and only the negative roots were used in the integration of (16). The items with b_g parameters maximizing (15) and (16) were "repeated" 10 times to yield 10 item RUC and conventional tests of the Type II and Type III variety. Average test information was computed and information functions plotted.

Before presenting the results for Phase I it might be helpful to summarize exactly what we produced in the way of tests and item banks. First we derived 25 ideal item banks. Each was defined by a value of a_g (.6, .8, 1.0, 1.5 or 2.0) and a guessing pattern (G_1, G_2, G_3, G_4 and G_5) -- the guessing patterns were necessary for RUC items but only the first value of each, c_{g1}, was necessary for binary items and served as the single guessing parameter. From each of the 25 ideal item banks seven varieties of tests were "constructed":

1) A series of perfect tailored tests yielding a constant and maximum test information function across θ. Each tailored test (one for each θ) was composed of 10 identical items.

2) A 10-item, Type I, RUC test composed of 10 identical items. The repeated item-type was chosen to maximize $\mu[I(\theta, Y_g)]$, the average RUC information function.

3) A 10-item, Type I, conventional test composed of 10 identical items (in terms of item parameters). The single, repeated item was chosen to maximize $\mu[I(\theta, U_g)]$, the average binary information function.

4) A 10-item, Type II, RUC Test composed of 10 identical items. The repeated item in the Type II RUC test was chosen to maximize $\mu[I(\theta, Y_g|\theta>0)]$, the average RUC information for average and above average ability.

5) A 10-item, Type II, Conventional test composed of 10 identical items. For this test, the repeated item was selected to maximize $\mu[I(\theta, U_g|\theta>0)]$, the conditional average binary information for average and above average ability.

6) A 10-item, Type III, RUC test composed of 10 identical items. The 10 identical items for this test were chosen so as to maximize $\mu[I(\theta, Y_g|\theta<0)]$, the conditional average RUC information for average and below average ability.

7) A 10-item, Type III, conventional test composed of 10 identical items. The repeated item here was that item maximizing $\mu[I(\theta, U_g|\theta<0)]$, the average binary information for average and below average ability.

Results for Phase I

Table 2 summarizes the average information yielded by perfect-tailored and Type I RUC and conventional tests. Since the information function for the perfect tailored tests (constructed from ideal item banks) is constant across ability, the ratio of either the average information in RUC or conventional tests to the average information for perfect tailored tests is equal to the average efficiency of RUC or conventional tests relative to perfect tailored tests. In Table 2 it is interesting to note that, with exception of guessing pattern G_5, Type I RUC tests were, on the average, essentially as efficient or more efficient than the perfect tailored tests for a_g = .6, .8 and 1.0. For the higher values of a_g (1.5 and 2.0), the Type I RUC tests did not fare well relative to tailored tests. For all 25 ideal item banks represented in Table 2, the Type I conventional tests compared

Table 2

Average Information for Perfect Tailored and Type I RUC and Conventional Tests

Guessing Pattern	a_g	Perfect Tailored Tests	Type I RUC Test	Type I Conventional Test	Average Relative Efficienty RUC	Average Relative Efficiency Conventional
G_1	.6	1.580	1.923	1.351	1.22	.86
	.8	2.809	3.084	2.179	1.10	.78
	1.0	4.390	4.329	3.070	.99	.70
	1.5	9.878	7.571	5.375	.77	.54
	2.0	17.560	10.944	7.697	.62	.44
G_2	.6	1.072	1.081	.904	1.01	.84
	.8	1.905	1.733	1.451	.91	.76
	1.0	2.977	2.430	2.036	.82	.68
	1.5	6.699	4.231	3.543	.63	.53
	2.0	11.908	6.04[illegible]	5.033	.51	.42
G_3	.6	1.89[illegible]	2.309	1.641	1.22	.86
	.8	3.372	3.706	2.660	1.10	.79
	1.0	5.268	5.206	3.763	.99	.71
	1.5	11.854	9.113	6.627	.77	.56
	2.0	21.074	13.189	9.544	.63	.45
G_4	.6	1.468	1.644	1.251	1.12	.85
	.8	2.610	2.637	2.014	1.01	.77
	1.0	4.077	3.700	2.834	.91	.70
	1.5	9.174	6.458	4.955	.70	.54
	2.0	16.309	9.288	7.082	.57	.43
G_5	.6	.857	.762	.720	.89	.84
	.8	1.523	1.220	1.153	.80	.76
	1.0	2.380	1.709	1.616	.72	.68
	1.5	5.355	2.967	2.807	.55	.52
	2.0	9.519	4.209	3.974	.44	.42

[1]Efficiencies are relative to the perfect tailored tests.

moderately well to the tailored tests only for the lowest value of a_g (.6). For a_g equal to one or greater, the Type I conventional tests were rather dismal in efficiency compared to the perfect tailored tests.

Summarizing the data comparing Tailored Tests and Type I tests, one can conclude that for values of a_g equal to one or less, RUC tests provide on the average, as much information as perfect tailored tests. Conventional Type I tests are about 85 percent as efficient, on the average, as tailored tests when a_g is as low as .6. For higher values of a_g, the conventional tests quickly drop in efficiency relative to tailored tests. Since averages can be misleading, it is useful to look at some of the actual information functions for a portion of the data. Figures 1, 2, 3, 4 and 5 depict the information functions for the perfect tailored tests and the Type I RUC and conventional tests for the five ideal item banks defined by G_1, G_2, G_3, G_4, G_5 and $a_g = 1$. In each figure, the circled line represents the perfect tailored tests, and the lines marked with triangles and x's represent, respectively, the RUC and conventional tests. The five figures conform to expectations formed by examining the average information values in Table 2; namely, that there is a range of abilities for which the information for the RUC test exceeds that of the tailored tests, and that there is a point where tailored and conventional tests are equal in efficiency. Additionally, notice in Figures 1 through 5 that both the Type I RUC and conventional tests are very inefficient compared to tailored tests at the extremes of the ability distribution. At the same time, however, there are very few examinees with extreme abilities -- assuming that ability is normally distributed. That the Type I RUC and conventional tests fared as well as they did compared to tailored tests, on the average, is due to the fact that their maximum information occurs in the region of the ability distribution where the probability densities are greatest.

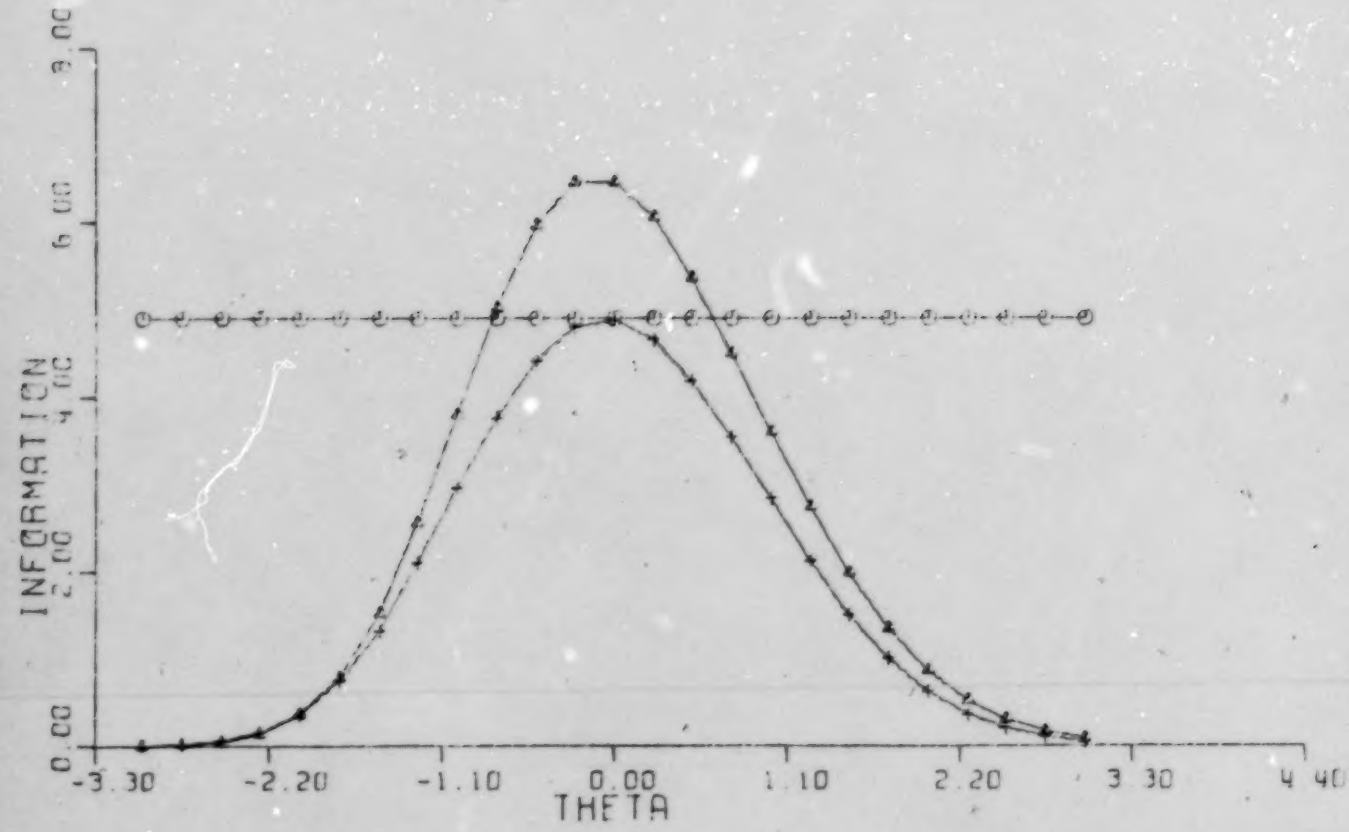

Figure 1. Information functions for perfect tailored and Type I RUC and conventional tests. Guessing pattern = G_1 and $a_g = 1$.

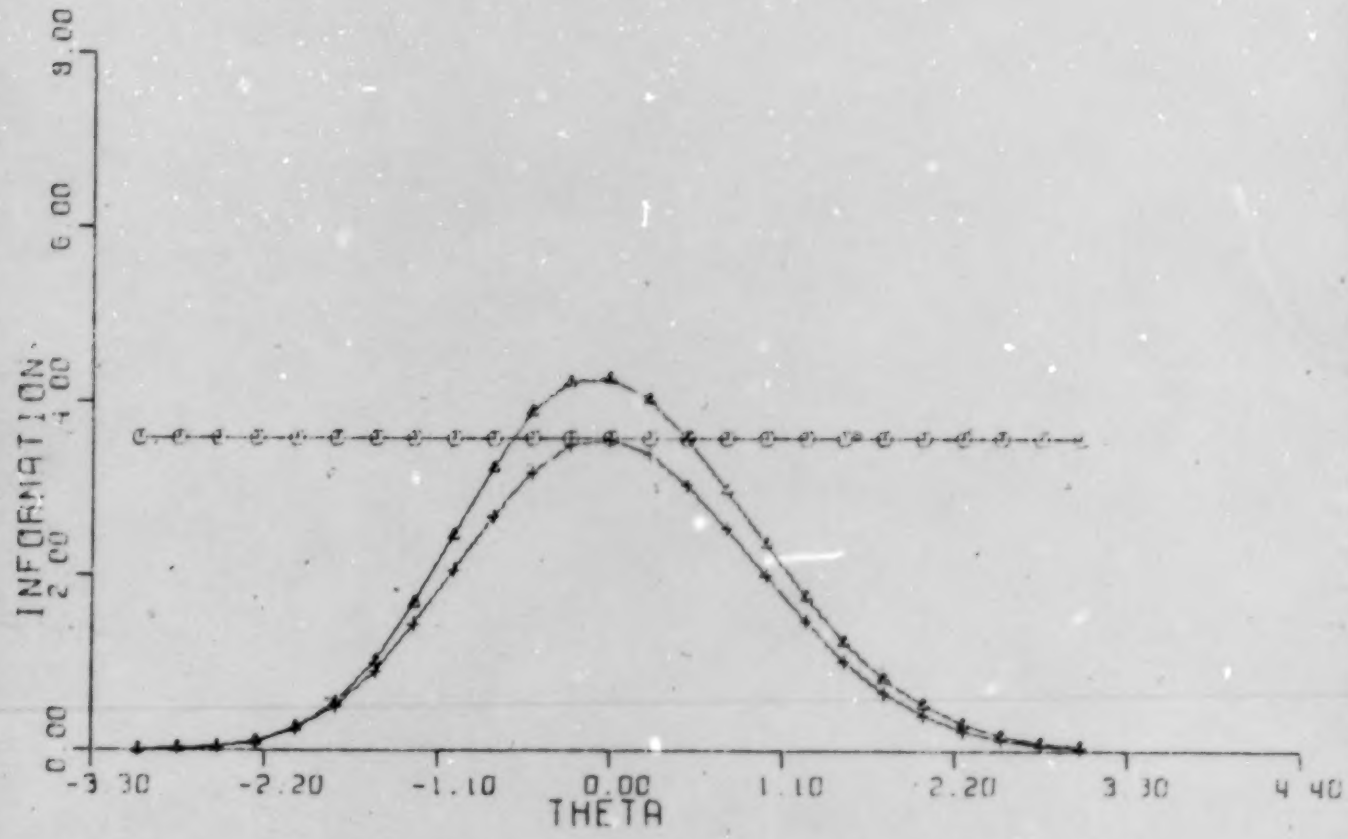

Figure 2. Information functions for perfect tailored and Type I RUC and conventional tests. Guessing pattern = G_2 and $a_g = 1$.

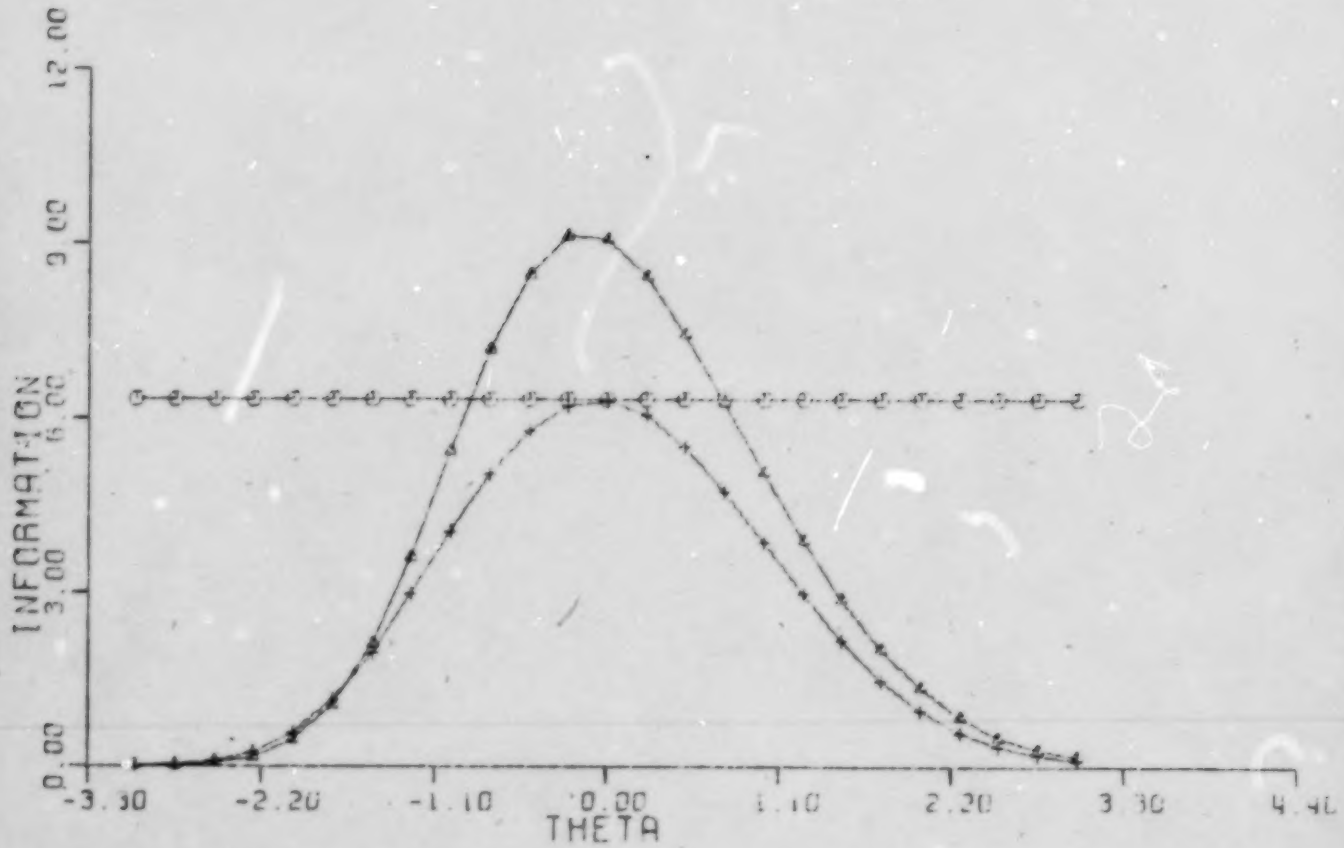

Figure 3. Information functions for perfect tailored and Type I RUC and conventional tests. Guessing pattern = G_3 and $a_g = 1$.

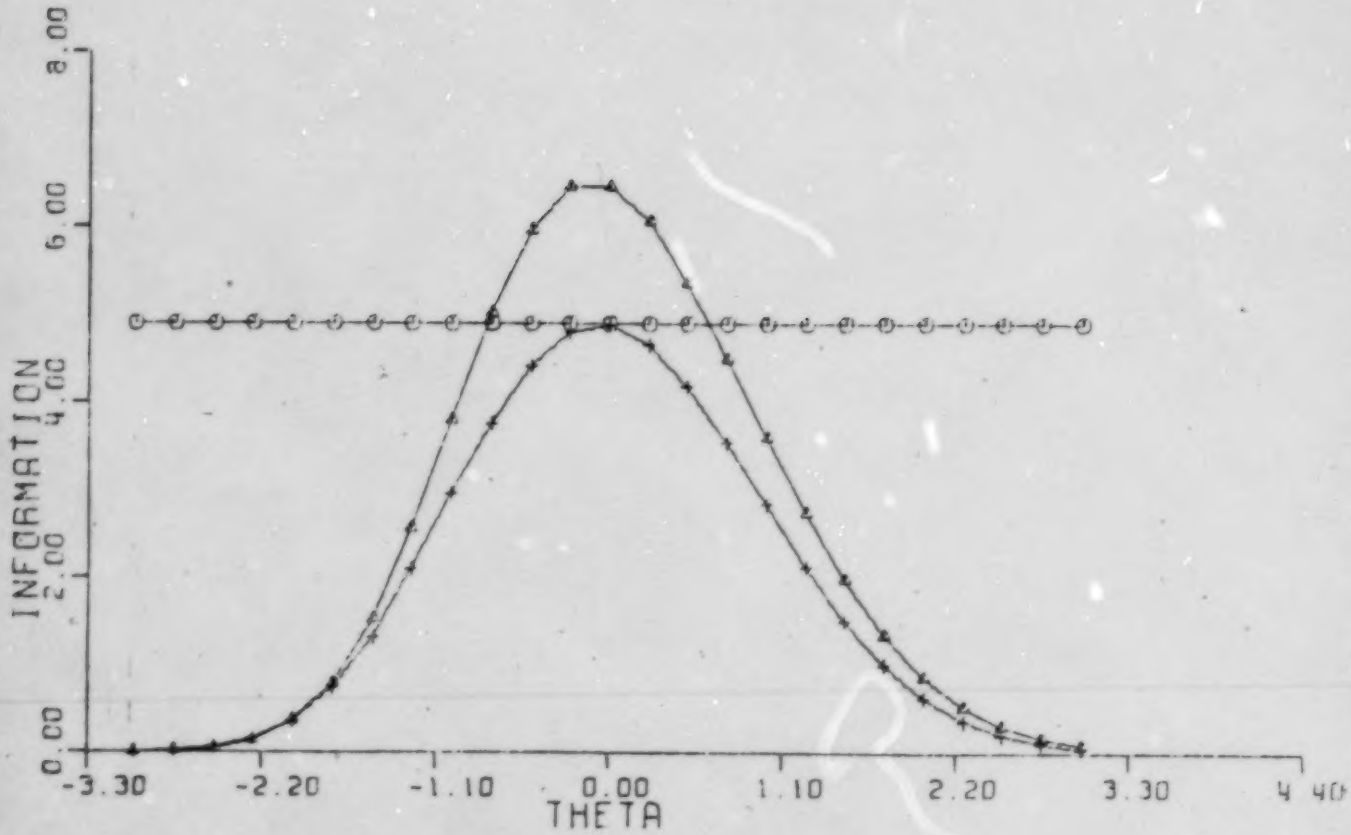

Figure 4. Information functions for perfect tailored and Type I RUC and conventional tests. Guessing pattern = G_4 and $a_g = 1$.

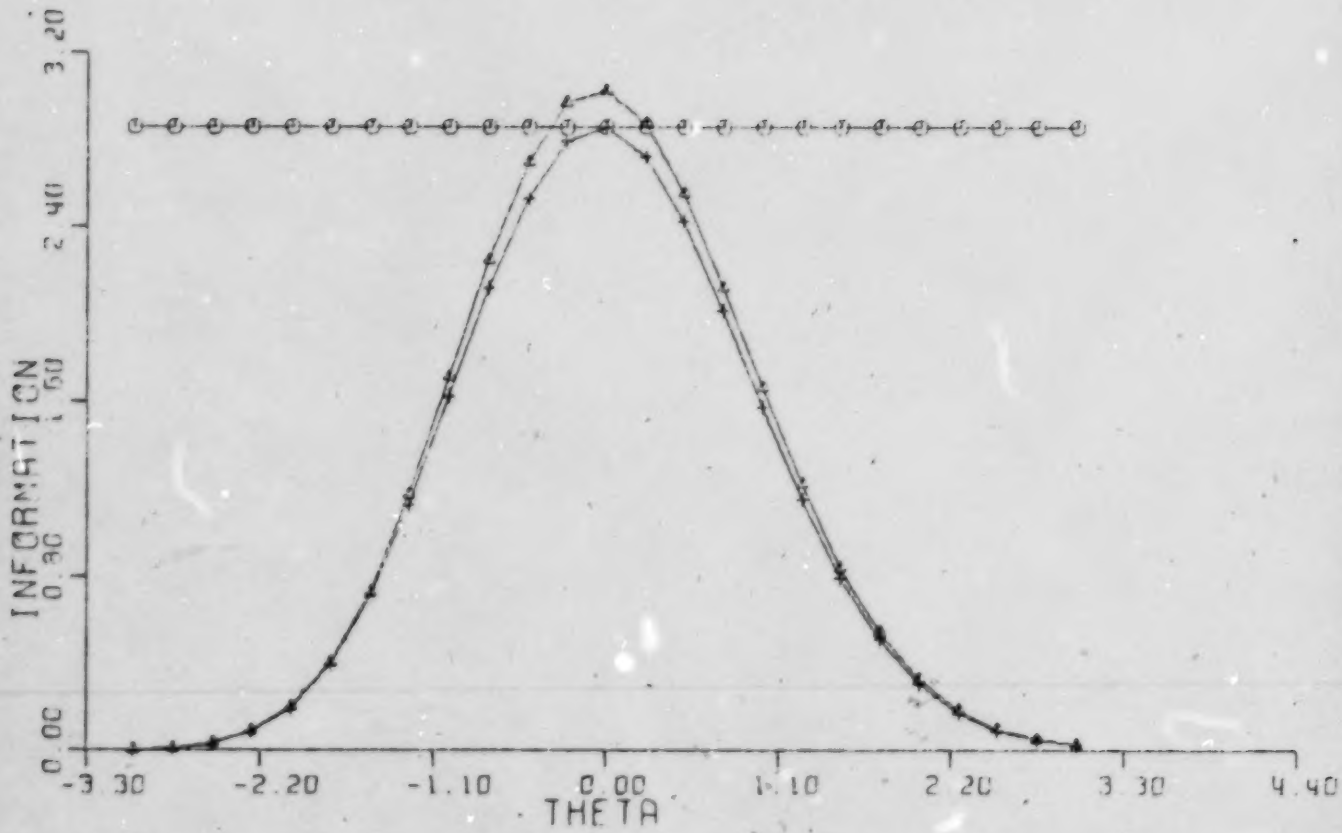

Figure 5. Information functions for perfect tailored and Type I RUC and conventional tests. Guessing pattern = G_5 and $a_g = 1$.

Since Figures 1-5 are for $a_g = 1$, it is important to discuss the effect of a_g on the information functions for the tailored, RUC and conventional tests. This effect may be succinctly summarized: as a_g increased the constant information in the tailored tests increased, and the range of abilities for which the RUC Tests yielded higher information than the perfect tailored tests and the conventional tests yielded information close to that of the perfect tailored tests narrowed considerably. This effect of a_g held for all guessing patterns and is obvious in Table 2 -- for any guessing pattern in Table 2, notice that the average relative efficiencies for both RUC and conventional tests decreased as a_g increased. When a_g was as low as .6, the RUC tests were more efficient, on the average, than tailored tests for all guessing patterns but G_5, and the conventional tests were approximately 85 percent as efficient, on the average, as the perfect tailored tests (across all guessing patterns).

For most selection situations in which a test is utilized, it is desireable to maximize information for examinees of high ability. The Type II RUC and conventional tests, constructed from the ideal item banks, were chosen so as to maximize average information for $\theta \geq 0$. The information functions for the perfect tailored tests, of course, remained the same as those used for comparison with the Type I fixed tests of the RUC and conventional variety.

Table 3 summarizes the average information obtained from the tailored and Type II RUC and conventional tests. Table 3 also lists the average efficiencies of RUC and conventional tests relative to perfect tailored tests. The Type II RUC tests for all ideal item banks defined by the guessing patterns G_1, G_3 and G_4 were higher in average information than the perfect tailored tests (with the exception of the RUC tests constructed from the ideal item bank defined by G_4 and $a_g = 2.0$). Additionally, for only three of the 25

Table 3

Average Information for Perfect Tailored and Type II RUC and Conventional Tests

Guessing Pattern	a_g	Perfect Tailored Tests	Type II RUC Test	Type II Conventional Test	Average[1] Relative Efficiency (RUC)	Average Relative Efficiency (Conventional)
G_1	.6	1.580	2.164	1.500	1.36	.95
	.8	2.809	3.709	2.574	1.32	.92
	1.0	4.390	5.562	3.863	1.27	.88
	1.5	9.878	12.360	7.782	1.25	.79
	2.0	17.560	18.123	12.431	1.03	.71
G_2	.6	1.072	1.215	1.012	1.13	.94
	.8	1.905	2.079	1.732	1.09	.91
	1.0	2.977	3.113	2.591	1.05	.87
	1.5	6.699	6.251	5.189	.93	.77
	2.0	11.908	10.014	8.274	.84	.70
G_3	.6	1.897	2.598	1.808	1.37	.95
	.8	3.372	4.454	3.110	1.32	.92
	1.0	5.268	6.683	4.679	1.27	.89
	1.5	11.854	13.511	9.473	1.14	.80
	2.0	21.074	21.814	15.158	1.04	.72
G_4	.6	1.468	1.848	1.392	1.26	.95
	.8	2.610	3.166	2.386	1.21	.91
	1.0	4.077	4.744	3.578	1.16	.88
	1.5	9.174	9.560	7.198	1.04	.78
	2.0	16.309	15.376	11.492	.94	.71
G_5	.6	.857	.856	.808	1.00	.94
	.8	1.523	1.463	1.381	.96	.91
	1.0	2.380	2.187	2.064	.92	.87
	1.5	5.355	4.375	4.123	.82	.77
	2.0	9.519	6.983	6.573	.73	.69

[1]Efficiencies are relative to perfect tailored tests.

ideal item banks, did the Type II RUC tests yield average information that was below 90 percent of the information for perfect tailored tests. The Type II conventional tests ranged from 69 to 95 percent as efficient as perfect tailored tests on the average. As was the case with Type I fixed tests, the Type II RUC and conventional tests were closer to (or superior to, in the case of RUC tests) perfect tailored tests in terms of average information when a_g was equal-to or less-than 1.0.

Figures 6 through 10 contain the actual information functions for the tailored and Type II RUC and conventional tests constructed from the ideal item banks defined by G_1, G_2, G_3, G_4 and G_5 and $a_g = 1.0$. In each graph the line marked with circles is the information function for perfect tailored tests; the line with triangles is for RUC tests, and the line marked with x's is for the conventional tests. By design, the Type II fixed tests (RUC and conventional) are poor providers of information for low and very high ability; whereas, the perfect tailored tests produced a uniformly high value of information across all abilities. Notice that the information functions in Figures 6-10 are virtually identical in shape to those given in Figures 1-5 with the exception that there is a shift in their location with respect to the ability axis. For Type I RUC and conventional tests, the maxima for the information functions occured in the vicinity of $\theta = 0$ (see Figure 1-5), and for Type II RUC and conventional tests, the maxima were in the vicinity of $\theta = .7$. Figures 6-10 do not reveal the effect of a_g on the information functions for the Type II tests. As was the case for Type I conventional and RUC tests, the range of abilities for which the RUC tests yielded higher information and the conventional tests provided information nearly equal to tailored tests, decreased monotonically with the magnitude of a_g. This relationship is obvious in Table 3 where the average efficiencies

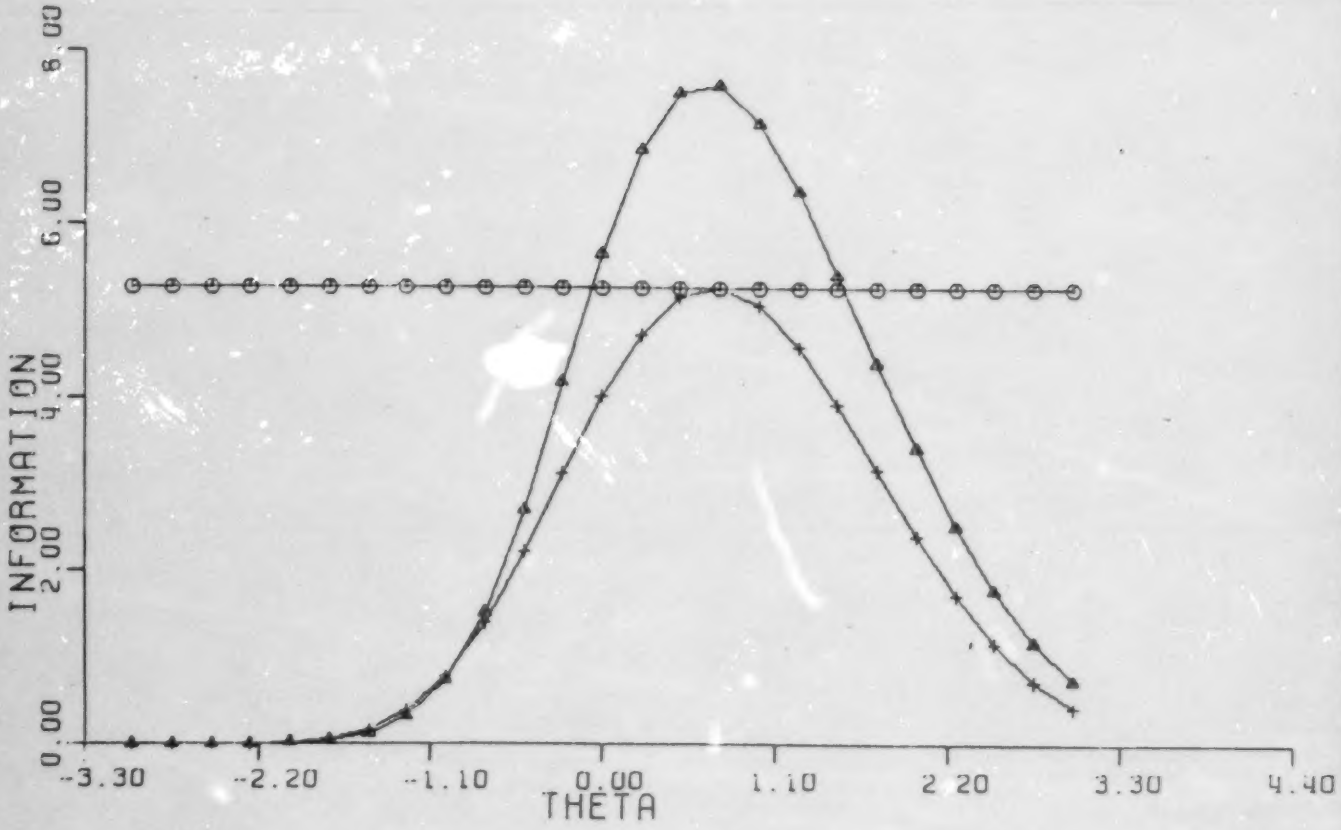

Figure 6. Information functions for perfect tailored and Type II RUC and conventional tests. Guessing pattern = G_1 and $a_g = 1$.

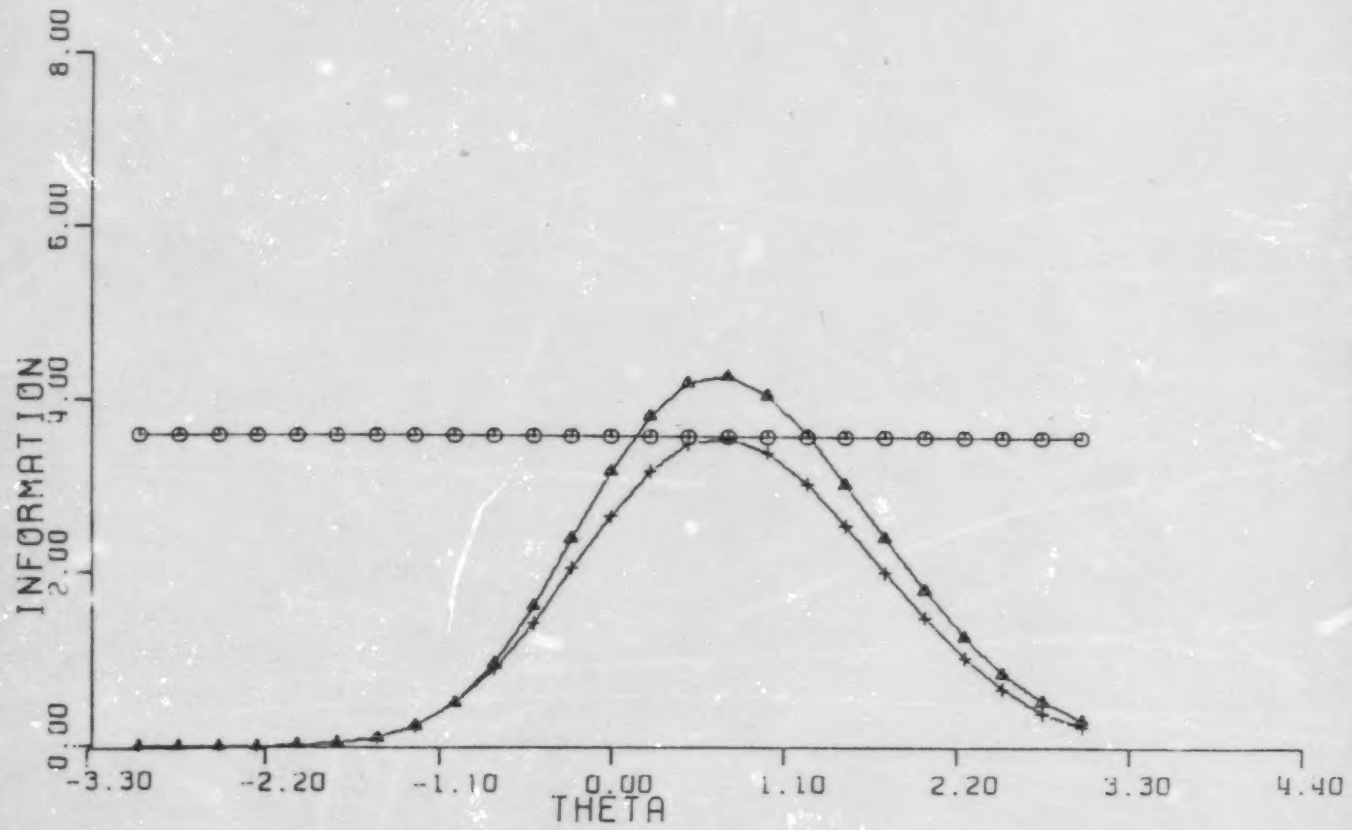

Figure 7. Information functions for perfect tailored and Type II RUC and conventional tests. Guessing pattern = G_2 and $a_g = 1$.

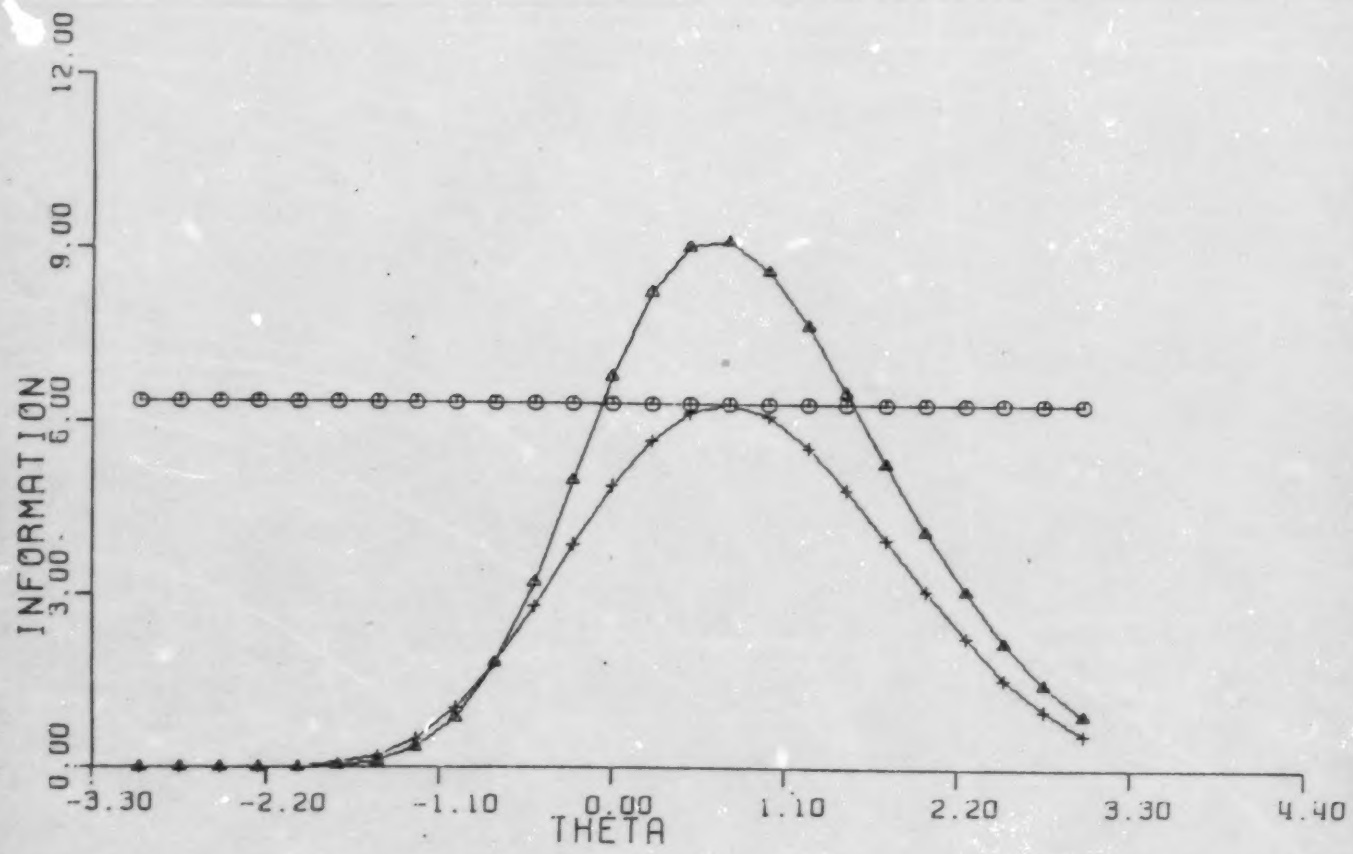

Figure 8. Information functions for perfect tailored and Type II RUC and conventional tests. Guessing pattern = G_3 and $a_g = 1$.

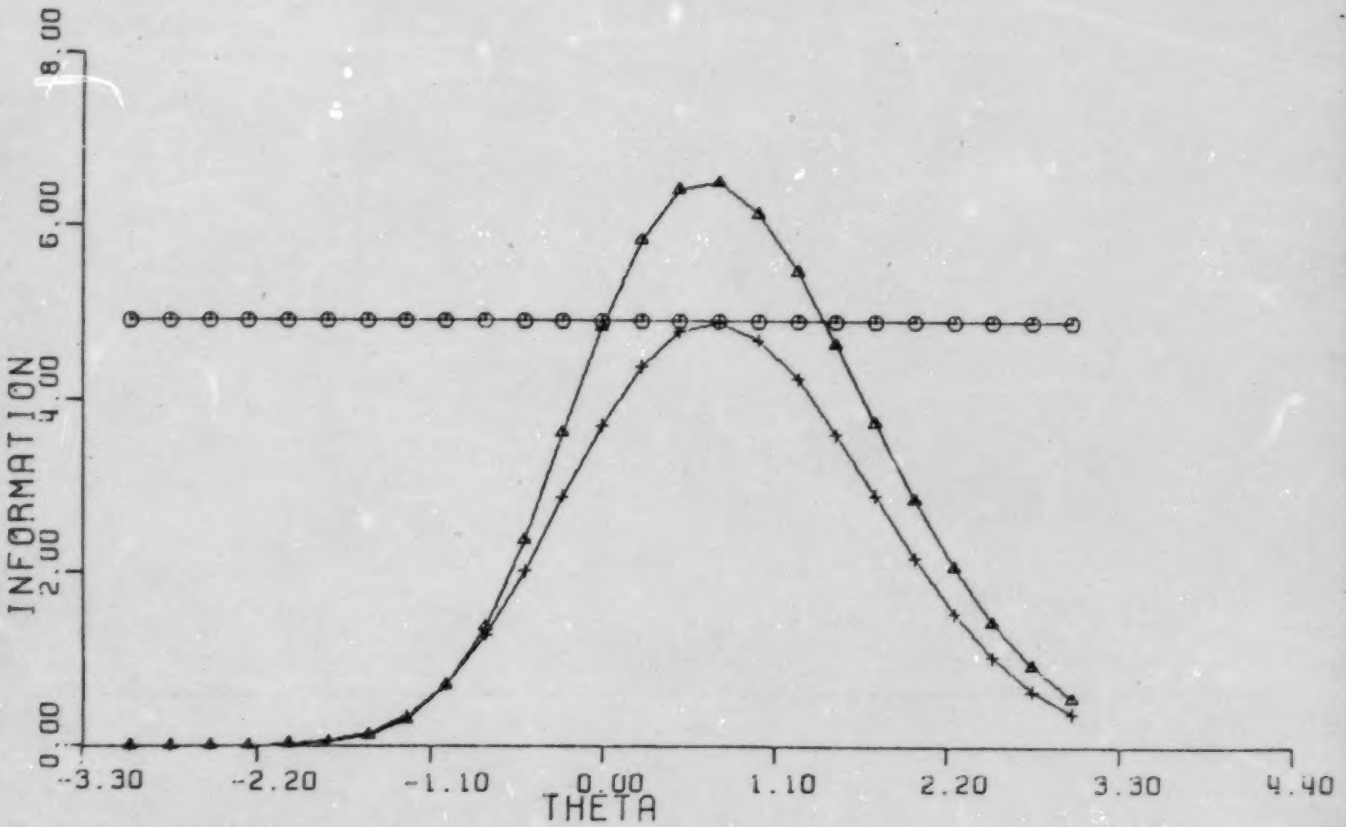

Figure 9. Information functions for perfect tailored and Type II RUC and conventional tests. Guessing pattern = G_4 and $a_g = 1$.

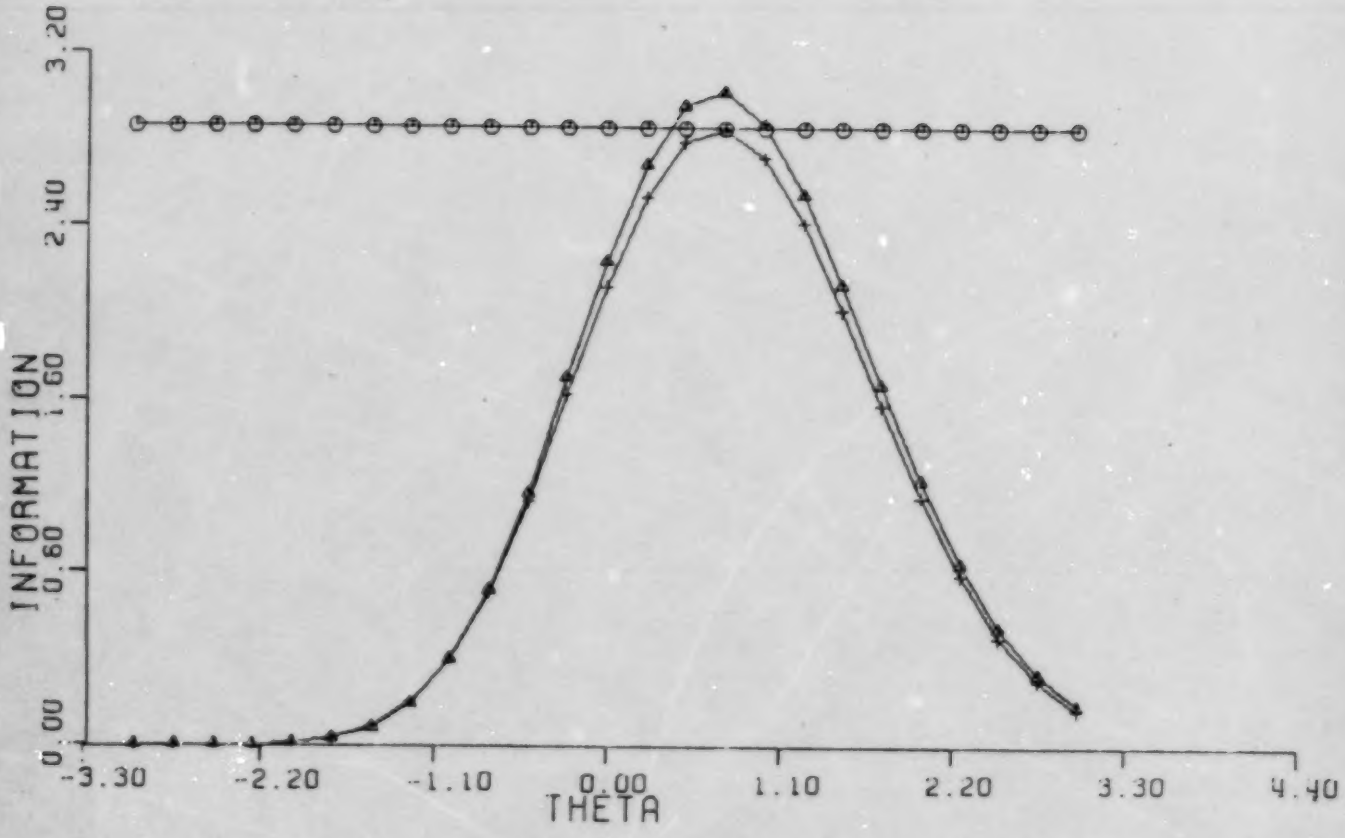

Figure 10. Information functions for perfect tailored and Type II RUC and conventional tests. Guessing pattern = G_5 and $a_g = 1$.

of Type II RUC and conventional tests (relative to tailored tests) decreased as a_g increased.

The last task of Phase I of the present inquiry was an examination of the functioning of Type III RUC and conventional tests (designed to be optimal for average and below average ability) relative to perfect tailored tests. The optimal ten-item conventional and RUC tests were constructed from the 25 ideal item banks by finding the value of b_g in each bank that maximized the conditional average information for $\theta \leq 0$ [see Equation (16)]. The 10-item perfect tailored tests were the same as those used for comparison with Type I and II fixed tests.

Table 4 presents the average information for the Type III conventional and RUC tests, as well as the average information for the tailored tests for the 25 ideal item banks. The last two columns of Table 4 list the average relative efficiencies of the RUC and conventional tests, where the demoninator of the efficiency ratios is the information for the perfect tailored tests. The information data given in Table 4 for the comparison of the two varieties of Type III fixed tests vs. perfect tailored tests is very similar to the corresponding data for the Type II tests presented in Table 3. The differences in the average information and average efficiencies for the Type II and Type III RUC and conventional tests is due to the greater influence of the guessing parameters in the lower half of the ability distribution. For the guessing patterns G_1, G_2, G_3 and G_4, and for a_g = .6, .8, and 1.0, the Type III RUC tests provided more information on the average than the perfect tailored tests (this outcome also held for the Type II RUC tests summarized in Table 3). The 15 ideal item banks defined by a_g = .6, .8 and 1.0 yielded conventional tests whose average efficiency, relative to tailored tests, was never less than 86 percent. Figures 11 through 15 display the

Table 4

Average Information for Perfect Tailored and Type III RUC and Conventional Tests

Guessing Pattern	a_g	Perfect Tailored Tests	Type III RUC Test	Type III Conventional Test	Average[1] Relative Efficiency (RUC)	Average Relative Efficiency (Conventional)
G_1	.6	1.580	2.142	1.495	1.35	.95
	.8	2.809	3.641	2.557	1.30	.91
	1.0	4.390	5.410	3.822	1.23	.87
	1.5	9.878	10.622	7.619	1.08	.77
	2.0	17.560	16.425	11.938	.94	.68
G_2	.6	1.072	1.207	1.009	1.13	.94
	.8	1.905	2.055	1.721	1.08	.90
	1.0	2.977	3.058	2.566	1.03	.86
	1.5	6.699	6.030	5.085	.90	.76
	2.0	11.908	9.401	7.973	.79	.67
G_3	.6	1.897	2.571	1.803	1.36	.95
	.8	3.372	4.371	3.092	1.30	.92
	1.0	5.268	6.496	4.634	1.23	.88
	1.5	11.854	12.758	9.294	1.08	.78
	2.0	21.074	19.708	14.632	.94	.69
G_4	.6	1.468	1.832	1.387	1.25	.94
	.8	2.610	3.118	2.370	1.20	.91
	1.0	4.077	4.637	3.540	1.14	.87
	1.5	9.174	9.124	7.046	.99	.77
	2.0	16.309	14.161	11.035	.87	.68
G_5	.6	.857	.852	.805	.99	.94
	.8	1.523	1.452	1.372	.95	.90
	1.0	2.380	2.162	2.045	.91	.86
	1.5	5.355	4.270	4.046	.80	.76
	2.0	9.519	6.694	6.353	.70	.67

[1]Efficiencies are relative to perfect tailored tests.

actual information functions for perfect tailored tests and the Type III fixed tests defined by $a_g = 1.0$. These graphs make clear that even though the RUC tests were often superior on the average to tailored tests in several of the ideal item banks, they provide very little information for abilities greater than $\theta = .5$ or less than $\theta = -1.5$. For values of a_g less than 1.0, the range of values of θ, for which the information for Type III fixed tests was either superior to (in the case of RUC tests) or nearly equal to (in the case of conventional tests) perfect tailored tests, increased. When a_g was set at values greater than 1.0, the range of θ values for which Type III RUC and conventional tests had greater or essentially equal information relative to perfect tailored tests, decreased.

Summary - Phase I

The idealized experiments accomplished in Phase I of this research made use of the conceptual entities labeled "item banks" and "perfect tailored tests". The utility of these conceptual entities should be fairly obvious by now: 1) The concept of perfect tailored tests allows one to examine the ultimate functioning of tailored tests unfettered by the "annoyance" and error introduced by the necessity of successively estimating ability during the course of testing, and 2) Ideal item banks allow the construction of perfect tailored tests with equal numbers of items that yield constant values of the test information function for all abilities; as a consequence, computation of average information for perfect tailored tests is trivial. The notion of an ideal item bank also allows one to directly determine RUC and conventional test items that are optimal in the sense of maximizing average test information -- this advantage, however, is minor compared to tailored tests with constant test information functions. That ideal item banks yield perfect tailored tests with constant information not only facilitates computation of average

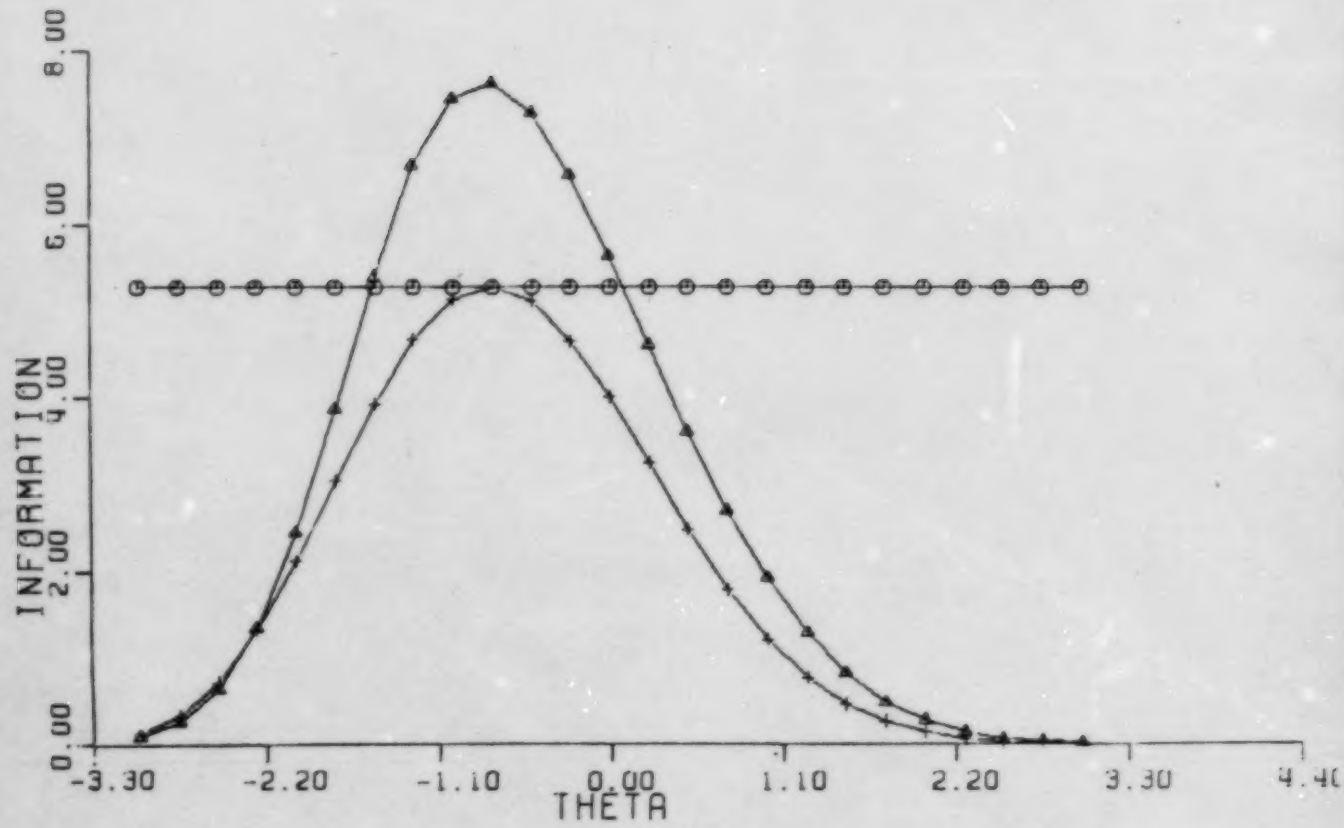

Figure 11. Information functions for perfect tailored and Type III RUC and conventional tests. Guessing pattern = G_1 and $a_g = 1$.

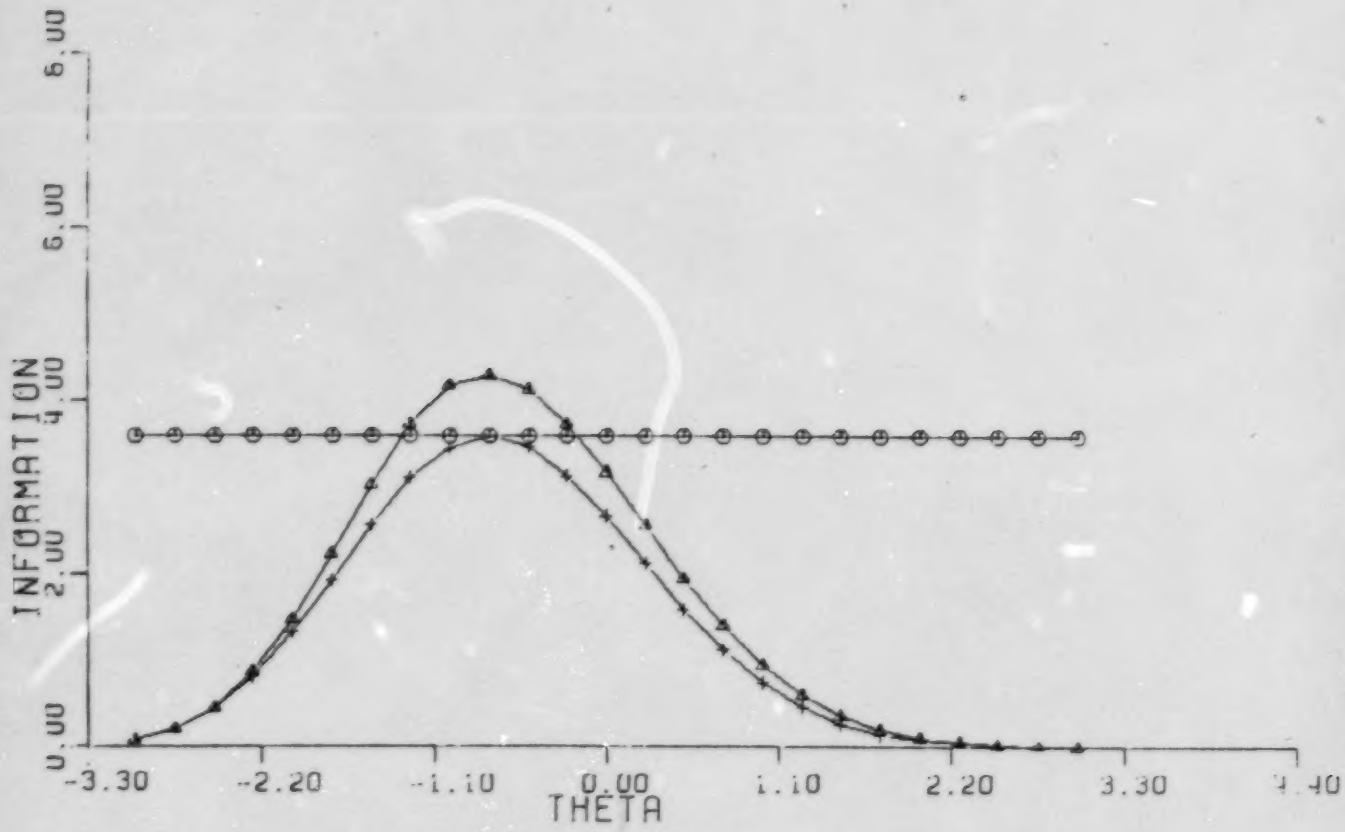

Figure 12. Information functions for perfect tailored and Type III RUC and conventional tests. Guessing pattern = G_2 and $a_g = 1$.

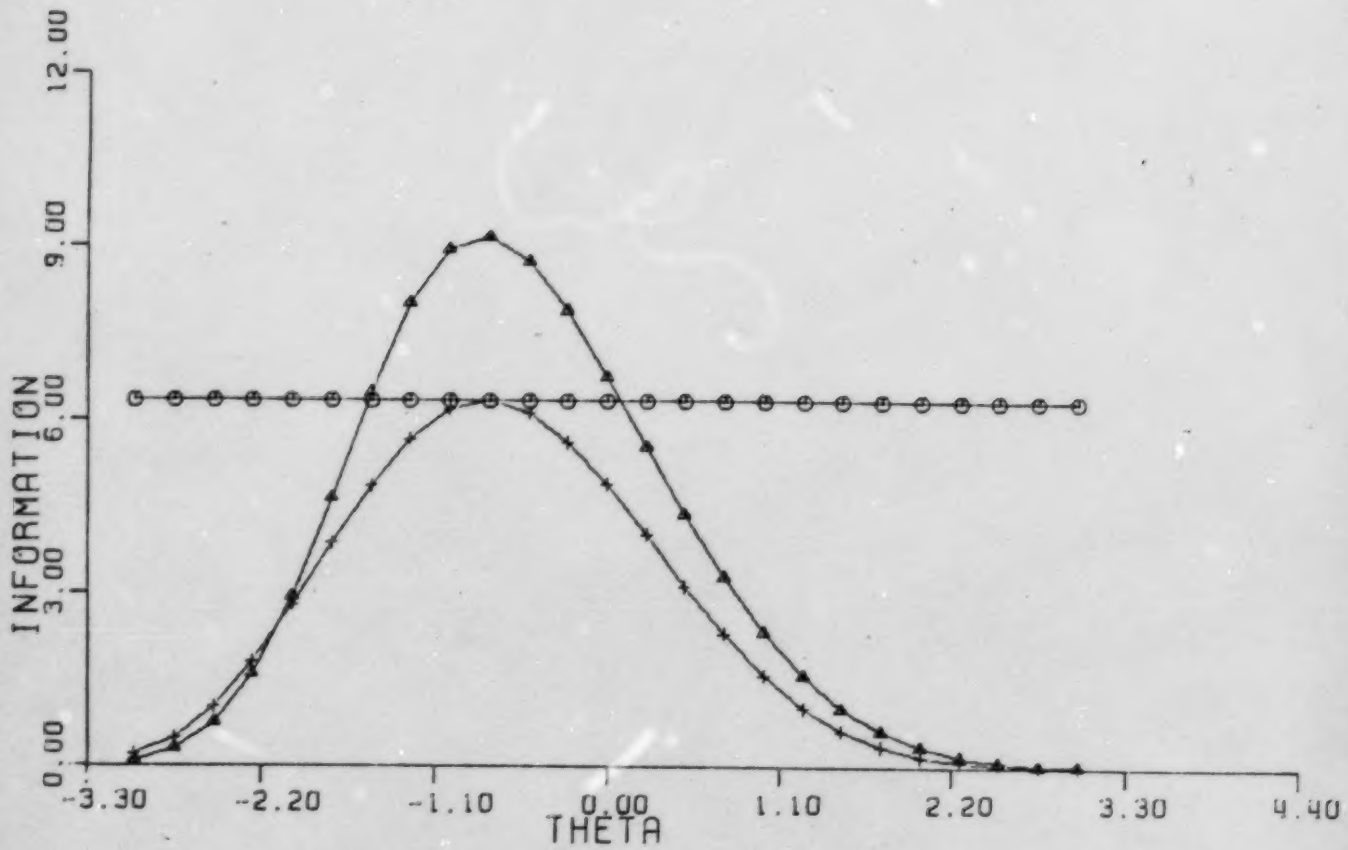

Figure 13. Information functions for perfect tailored and Type III RUC and conventional tests. Guessing pattern = G_3 and $a_g = 1$.

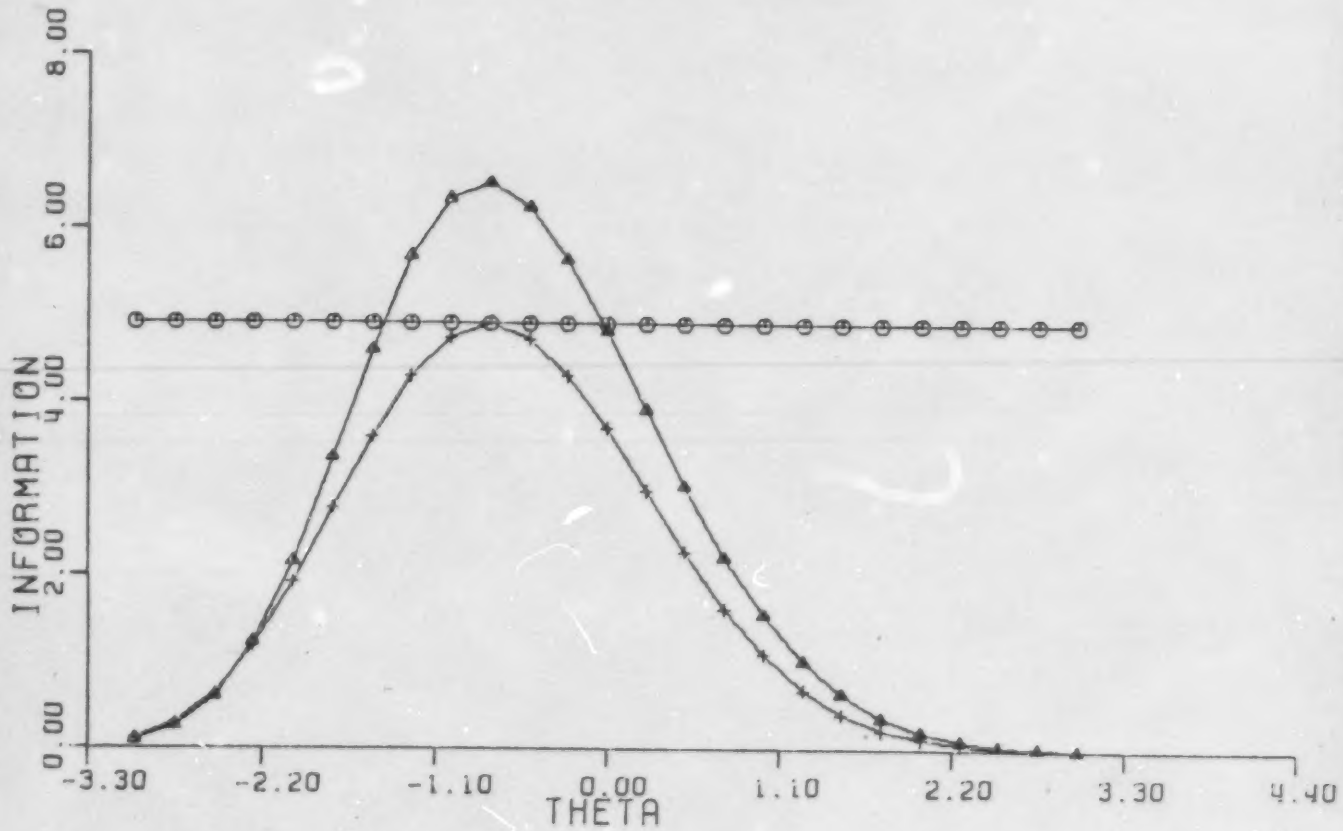

Figure 14. Information functions for perfect tailored and Type III RUC and conventional tests. Guessing pattern = G_4 and $a_g = 1$.

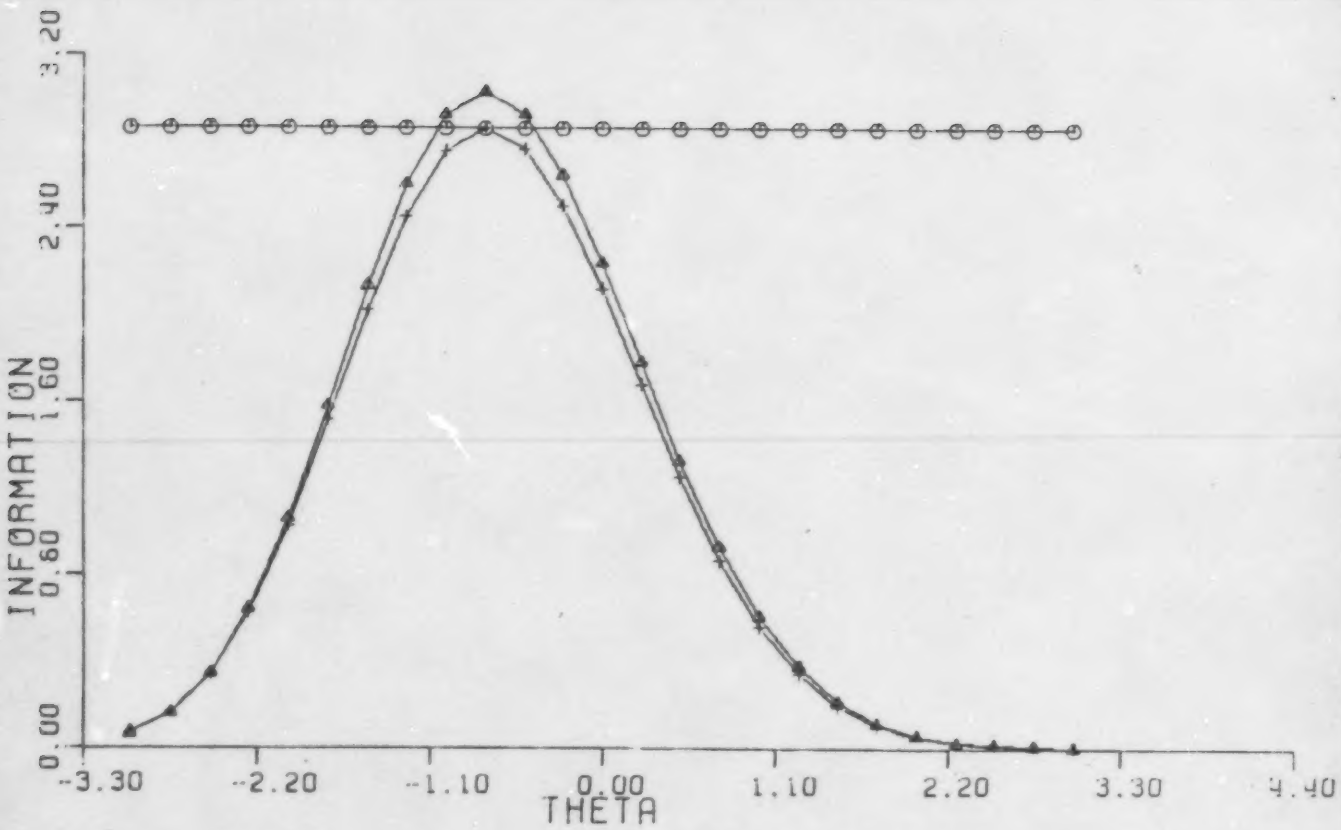

Figure 15. Information functions for perfect tailored and Type III RUC and conventional tests. Guessing pattern = G_5 and $a_g = 1$.

information across the ability distribution, but also allows one to compute the average efficiency of fixed RUC and conventional tests (relative to perfect tailored tests) by taking the ratio of the average information for fixed tests to the average (constant) information for tailored tests.

A summary of the results of the idealized experiments is given below. All statements concerning the RUC tests are conditional on the truth of the derived model for these types of items given in Equation (9). This model for RUC scores has not been confronted with actual data; however, the model is quite general and flexible (due to the large number of free parameters), and should be a fairly close approximation to reality given that the Knowledge or Random Guessing principle is incorrect.

1) Type I RUC tests (those RUC tests designed to maximize average information across the entire ability distribution) produced as high or higher average information than perfect tailored tests in seven out of 25 ideal item banks. In 12 of the 25 ideal item banks, the RUC test's average information was at least 90 percent that of the perfect tailored tests. Type I conventional tests did not perform well relative to perfect tailored tests; at best their average information was 86 percent that of tailored tests (see Table 2).

2) Type II fixed tests (those RUC and conventional tests constructed to maximize average information for $\theta \geq 0$) fared quite well, on the average, compared to perfect tailored tests. In 18 out of the 25 ideal item banks, the Type II RUC tests produced average information greater-than or equal-to the average information for perfect tailored tests, and in 22 of the 25 ideal item banks, the Type II RUC tests were at least 90 percent as efficient on the average relative to tailored tests. For all 10 ideal banks defined by $a_g = .6$ or .8, the conventional tests were at least 90 percent as efficient, on the average, relative to perfect tailored tests. The conventional tests drawn from the 10 ideal item banks defined by $a_g = 1.5$ or 2.0 were only about 75 percent as efficient as the corresponding tailored tests (see Table 3).

3) Type III RUC and conventional tests were constructed from the 25 ideal item such that average information was maximized for $\theta < 0$; i.e., the Type III tests composed of fixed sets of items (either RUC or binary) were designed to maximize average information in the lower one-half of the ability distribution. Comparison of the Type III RUC tests with perfect tailored tests

indicated that for 14 of the 25 ideal item banks, the RUC tests had greater average efficiency than the perfect tailored tests, and in 21 of the 25 ideal item banks, the RUC tests were at least 90 percent as efficient as the perfect tailored tests. The 10 conventional Type III tests drawn from the ideal item banks defined by a_g = .6 or .8, were at least 90 percent as efficient as perfect tailored tests on the average. When the a_g values defining an ideal item bank were equal to 1.5 or 2.0, the Type III conventional tests ranged from 67 to 88 percent as efficient as perfect tailored tests (see Table 4).

In concluding the discussion of Phase I, we feel that it has been demonstrated that interactive testing, using a fixed set of RUC items for all examinees, shows promise for being as effective or more effective than tailored testing in certain situations; e.g., for examinees of high or low ability. Our demonstration, however, is based on idealized situations that do not exist in reality, and furthermore, is based on a summary index -- the average test information function. There is no doubt that some psychometricians will find repellent the notion of averaging information functions, because an average test information function can disguise the fact that a test is measuring poorly in some portions of the ability distribution. The fact remains that if an average efficiency for a RUC test, relative to tailored tests, exceeds unity, it is clear that for most of the ability distribution, the RUC test is providing more information than tailored tests. On the other hand, Figures 1-15 indicate that even when RUC tests are more effective than tailored tests on the average, there is a small proportion of the ability distribution poorly assessed by the RUC tests compared to tailored tests.

A last word should be inserted here in defense of conventional tests consisting of a fixed-set of binary items. Tailored and RUC tests probably take one-and-a-half to two times as long to administer as a conventional test of equal length. If the conventional tests studied here were all increased one-and-a-half times in length (i.e., contained 15 instead of 10 items), then in

every instance they would be more efficient on the average than RUC tests, and in most instances would yield greater average information than perfect tailored tests.

Phase II. A Comparison of Perfect Tailored Tests, RUC Tests and Conventional Tests using Item Banks of Finite Size.

Phase II of this research project was identical to Phase I with the exception that the ideal item banks of Phase I were replaced with two item banks of finite length -- each containing items with a fairly broad range of values for the item parameters a_g and b_g. An item bank containing 200 calibrated items due to Urry (1974) and a collection of 72 calibrated vocabulary items compiled by Koch and Reckase (1978) were used in this phase of the research. In the first part of Phase II, the actual values of the item's c_g estimates were ignored and replaced by the five guessing patterns used in Phase I. More explicitly, for the binary items used in the tailored and conventional tests, the c_g values were taken to be equal to the "first guess" probabilities (c_{g1}) from the five guessing patterns described earlier, and, as a consequence, the values of c_g for binary items were either .10, .20, .24, .40 or .50 (the values of c_{g_1} from the guessing patterns G_3, G_1, G_4, G_2 and G_5 respectively). For RUC items the complete guessing patterns were used.

For this phase of the research one can think of the method as involving the construction and comparison of perfect tailored tests and Types I, II and III conventional and RUC tests from 10 different item banks of finite length. Five item-banks, consisting of 200 items each, were devised using the a_g and b_g values from Urry's items in combination with each of five guessing patterns; in other words, Urry's 200 items, each having guessing pattern G_1 imposed on them, constituted one item bank. Another item bank containing 200 items was

generated using Urry's items and guessing pattern G_2, and so forth. Likewise, the 72 items calibrated by Koch and Reckase were used to generate five item banks (each containing 72 items) by using the actual a_g and b_g values for the original items in combination with each of five quessing patterns.

When using item banks of finite length, as opposed to ideal item banks, the mathematical tractibility in computing the average information for perfect tailored tests is lost; however, perfect tailored tests can certainly be constructed from item banks of finite length, but obviously not for all values of θ. In this phase of the investigation, a discrete approximation to the continuous random variable, θ, was used. Let θ_i represent a discrete value of ability and $p(\theta_i)$ be its (marginal) probability. For each of 49 discrete values of θ_i ranging from -3 to +3 in equal steps of .125, the values $p(\theta_i)$ were computed as

$$(17) \qquad p(\theta_i) = \int_{\theta_i - .125}^{\theta_i + .125} \phi(t)dt \,, \qquad (i = 1, 2, \ldots, 49)$$

where $\phi(t)$ is the normal density function. Collectively these 49 values of θ_i and their corresponding $p(\theta_i)$ determined a discrete approximation to a normally distributed θ with mean zero and variance one.

The 10 perfect tailored tests for the five finite item-banks generated from Urry's 200 calibrated items, and the five finite item banks derived from Koch and Reckase's 72 items were constructed as follows:

1) For each of the 49 values of θ_i, ranging from -3 to +3 in steps of .125, the information was computed for each item in the particular bank. Those items having the highest information at each discrete ability value comprised the perfect tailored test for that ability.

2) The number of items in each perfect tailored test was equal to the smaller of two numbers -- a) the number of items that produced a test information value of 20 or more, or b) 25.

Applying this rule to the item banks derived from Urry's 200 items gave perfect tailored tests of 10 to 25 items, and for the Koch and Reckase item banks, the number of items in the perfect tailored tests ranged from 14 to 25.

The average number of items, $\bar{n}_I$, in each of the 10 sets of perfect tailored tests for all ability values (one set for each of 10 item banks), was computed as

$$(18) \qquad \bar{n}_I = \sum_i n_i \, p(\theta_i) \, , \qquad (i = 1,2,\ldots,49)$$

where n_i is the number of items in the perfect tailored test for the i^{th} value of θ. The value $\bar{n}_{II}$, the average number of items in the perfect tailored tests for above average ability ($\theta_i \geq 0$), was computed in each item bank using,

$$(19) \qquad \bar{n}_{II} = \sum_i n_i \, p(\theta_i \mid \theta_i \geq 0) \qquad (i = 1,2,\ldots,49) \, ,$$

and $\bar{n}_{III}$, the average number of items in the perfect tailored tests for below average ability ($\theta_i \leq 0$), was computed as

$$(20) \qquad \bar{n}_{III} = \sum_i n_i \, p(\theta_i \mid \theta_i \leq 0) \, . \qquad (i = 1,2,\ldots,49)$$

These average numbers of items (rounded to the nearest integer value) in the perfect tailored tests were used to determine the number of items used in the Type I, II and III conventional and RUC tests. Finally, average test information for each of the 10 sets of perfect tailored tests was computed -- the computational details are discussed later.

The optimal Type I, II and III RUC and conventional tests for each of the 10 finite item banks were determined as follows:

1) For Type I RUC and conventional tests, the number of items in each test was determined by the average number of items, $\bar{n}$, in the corresponding perfect tailored tests ($\bar{n}$ was set equal to the nearest integer value). The optimal RUC and conventional tests for each item bank consisted of the $\bar{n}$ items yielding the highest average information across the entire ability distribution. The average test information for the $\bar{n}$-item tests was computed by summing the average information across items.

2) The Type II and III conventional and RUC tests were constructed by choosing the $\bar{n}$ items yielding the highest conditional average information for $0<\theta\leq 3$ and $-3\leq\theta<0$ respectively. The average test information for these types of tests was found by summing the average item information across the $\bar{n}$ items (i.e., the expected value of a sum is equal to the sum of the expected values).

The average test information for the tailored and Type I RUC and conventional tests was found by the discrete approximation,

$$(21) \qquad \hat{\mu}[I(\theta)] = \sum_g \sum_i I(\theta_i, S_g)\, p(\theta_i) , \quad (-3<\theta\leq 3)$$

where S_g represents either the binary item score, U_g (for tailored and conventional tests) or the RUC score, Y_g. The average test information for Types II and III conventional and RUC tests was computed using,

$$(22) \qquad \hat{\mu}[I(\theta|\theta\geq 0)] = \sum_g \sum_i I(\theta_i, S_g)\, p(\theta_i|\theta_i\geq 0) ,$$

for Type II tests, and

$$(23) \qquad \hat{\mu}[I(\theta|\theta\leq 0)] = \sum_g \sum_i I(\theta_i, S_g)\, p(\theta_i|\theta_i\leq 0) ,$$

for Type III tests, where S_g represents either the binary or RUC score for the g^{th} item.

Results for Phase II

Table 5 summarizes the comparisons (in terms of average test information) of perfect tailored tests and the Types I, II and III RUC and conventional tests constructed from the five item-banks (one for each guessing pattern) derived from Urry's 200 calibrated items. The first three columns of Table 5 contain the average information values based on a discrete approximation to the continuous ability parameter, θ. The last two columns of Table 5, labeled "Efficiency Ratios", are the ratios of average information in the RUC or conventional tests to the average information in the perfect tailored tests. These ratios are not average relative efficiencies as their counterparts were in

Phase I (the test information for the perfect tailored tests used here is not constant across ability). Since the ratio of expected values is greater-than or equal-to the expected value of a ratio, the efficiency ratios in Table 5 are upper-bounds to the average relative efficiencies. One possible point of confusion in Table 5 (and the one to follow) should be clarified. The guessing pattern G_1, for example, used in conjunction with Urry's a_g and b_g values for 200 items, generated a single set of perfect tailored tests (one test for each of 49 values of θ_i). For the full ability distribution for the discrete θ_i, the perfect tailored tests gave average information equal to 18.9480 and contained an average of $\bar{n}_I = 21$ items; however, for above and below average ability, the (conditional) average test information functions were 20.2347 and 17.8020, respectively, and the (conditional) average numbers of items ($\bar{n}_{II}$ and $\bar{n}_{III}$) in the tailored tests were 18 (for above average ability) and 25 (for below average ability). The idea here was to choose a single set of perfect tailored tests for each finite item bank (defined by a particular guessing pattern), and then to average the test information and numbers of items in the tailored tests across various portions of the ability distribution.

Table 5 indicates that the average information for Type I RUC tests (containing from 17 to 25 items, depending on the guessing pattern) ranged from 78 to 95 percent of the average information for perfect tailored tests. Type I conventional tests compared poorly with perfect tailored tests; their average information was only about 65 percent that of perfect tailored tests. The Type II RUC tests contained between 19 and 25 items, and for all but G_5, gave higher conditional average information than did the perfect tailored tests. The Type II conventional tests (containing the same numbers of items as the

Table 5

Average Information and Efficiency Ratios
for Perfect Tailored Tests, RUC Tests and Standard Tests.
Item Banks based on Urry (1974).

Guessing Pattern	Perfect Tailored Tests (Avg. no. items)		RUC Tests	Conventional Tests	Efficiency[1] Ratio (RUC Tests)	Efficiency Ratio (Conventional Tests)
(1) For full ability distribution (Type I RUC and Conventional Tests)						
G_1	18.9480	(21)	17.9527	12.4819	0.95	0.66
G_2	14.3889	(25)	11.1906	9.3246	0.78	0.65
G_3	19.8033	(17)	18.4986	13.2664	0.93	0.67
G_4	18.2213	(22)	15.7372	11.8819	0.86	0.65
G_5	11.3768	(25)	7.7915	7.3519	0.68	0.65
(2) For above average ability (Type II RUC and Conventional Tests)						
G_1	20.2347	(18)	24.3886	17.2236	1.21	0.85
G_2	17.1032	(25)	17.1639	14.2883	1.00	0.84
G_3	20.5231	(14)	24.4312	17.6670	1.19	0.86
G_4	20.0973	(20)	22.4916	17.1470	1.12	0.85
G_5	13.5693	(25)	11.9739	11.2884	0.88	0.83
(3) For below average ability (Type III RUC and Conventional Tests)						
G_1	17.8020	(25)	22.4666	15.7475	1.26	0.88
G_2	11.6706	(25)	12.4231	10.3311	1.06	0.89
G_3	19.1687	(20)	22.9176	16.4475	1.20	0.86
G_4	16.5604	(25)	19.1099	14.4958	1.15	0.88
G_5	9.1731	(25)	8.6604	8.1646	0.94	0.89

[1]Ratio of average information in tests with fixed items (RUC or Conventional) to average information in tailored tests.

corresponding RUC tests) were about 85 percent as efficacious as perfect tailored tests in terms of conditional average information. The Type III RUC tests gave average information higher than the perfect tailored tests constructed from Urry's items with the exception of the item bank defined by G_5. The Type III conventional tests for the five item banks generated from Urry's data gave about 88 percent of the average information in the corresponding perfect tailored tests. It is interesting to note in Table 5 that the efficiency ratios for Type I, II and III conventional tests were virtually unaffected by the values of c_g which ranged between .10 (for pattern G_3) and .5 (for pattern G_5).

The actual test information functions for the Type I, II and III RUC and conventional tests and perfect tailored tests, drawn from the item bank defined by imposing pattern G_1 on Urry's item data, are presented in Figures 16, 17 and 18. In each of these figures the information curve marked with triangles is for RUC tests; the one marked with circles is for conventional tests, and the curve marked with x's is for perfect tailored tests. In these three figures, notice that there was a rather broad range of abilities for which the RUC tests outperformed the perfect tailored tests, but that the RUC tests were quite inferior to the tailored tests at the extremes of the ability continuum. There are, however, very few examinees at the extremes of the ability continuum assuming a normal distribution of ability, and for this reason, the Type II and III RUC tests gave higher average information than perfect tailored tests for the item bank defined by a combination of guessing pattern G_1 and Urry's items. It may seem peculiar that the information functions for the conventional Type I and II tests, depicted in Figures 16 and 17, exceeded the corresponding function for the perfect tailored tests

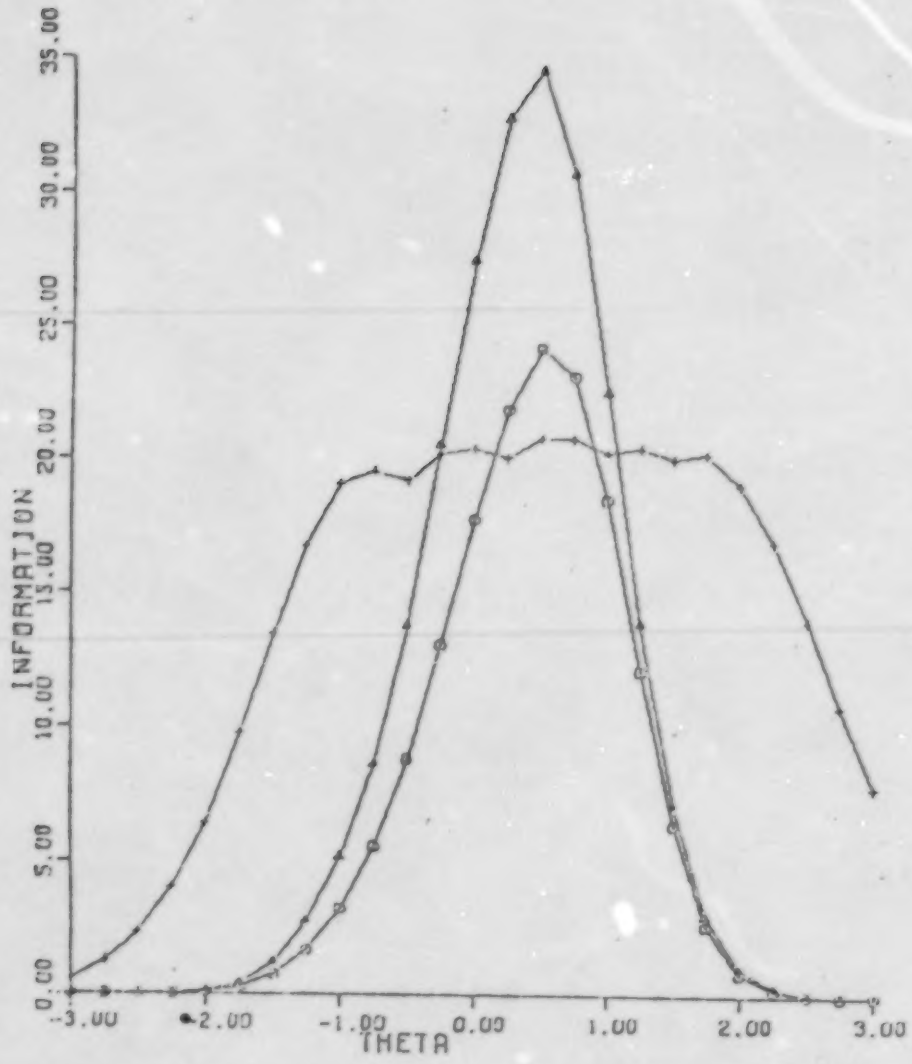

Figure 16. Information functions for perfect tailored and Type I RUC and conventional tests for the item bank defined by Urry's 200 items and guessing pattern G_1 .

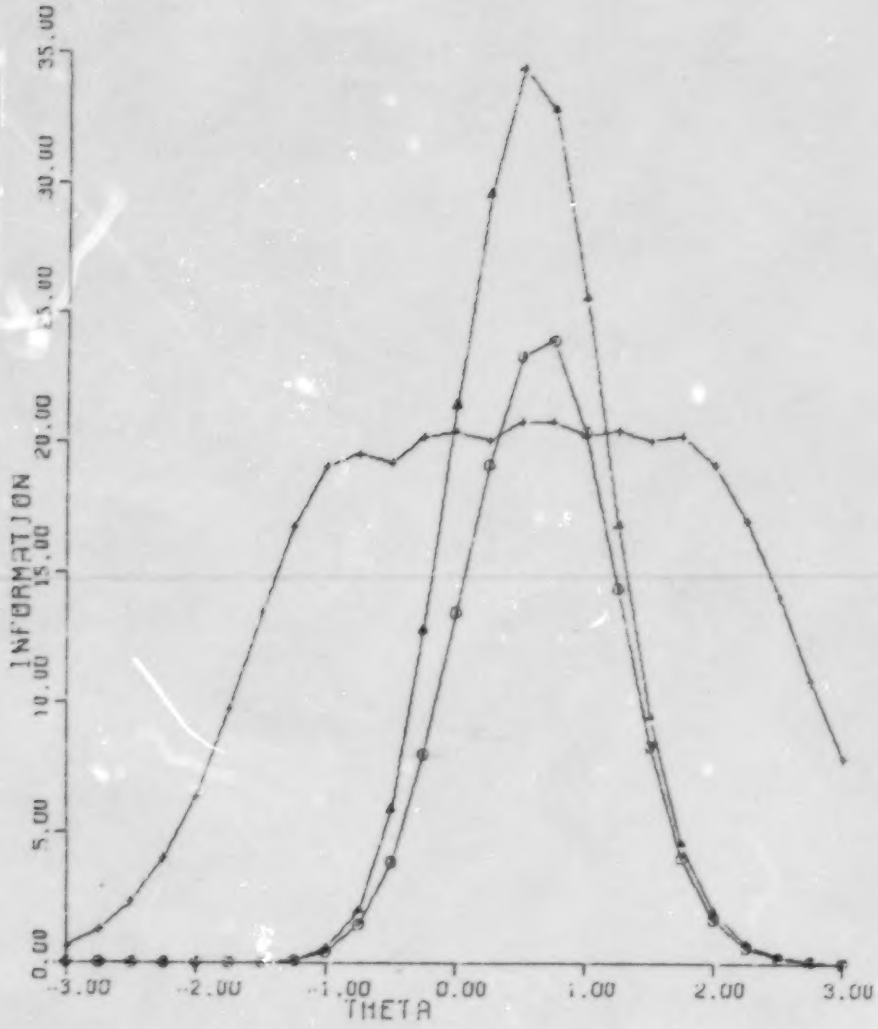

Figure 17. Information functions for perfect tailored and Type II RUC and conventional tests for the item bank defined by Urry's 200 items and guessing pattern G_1 .

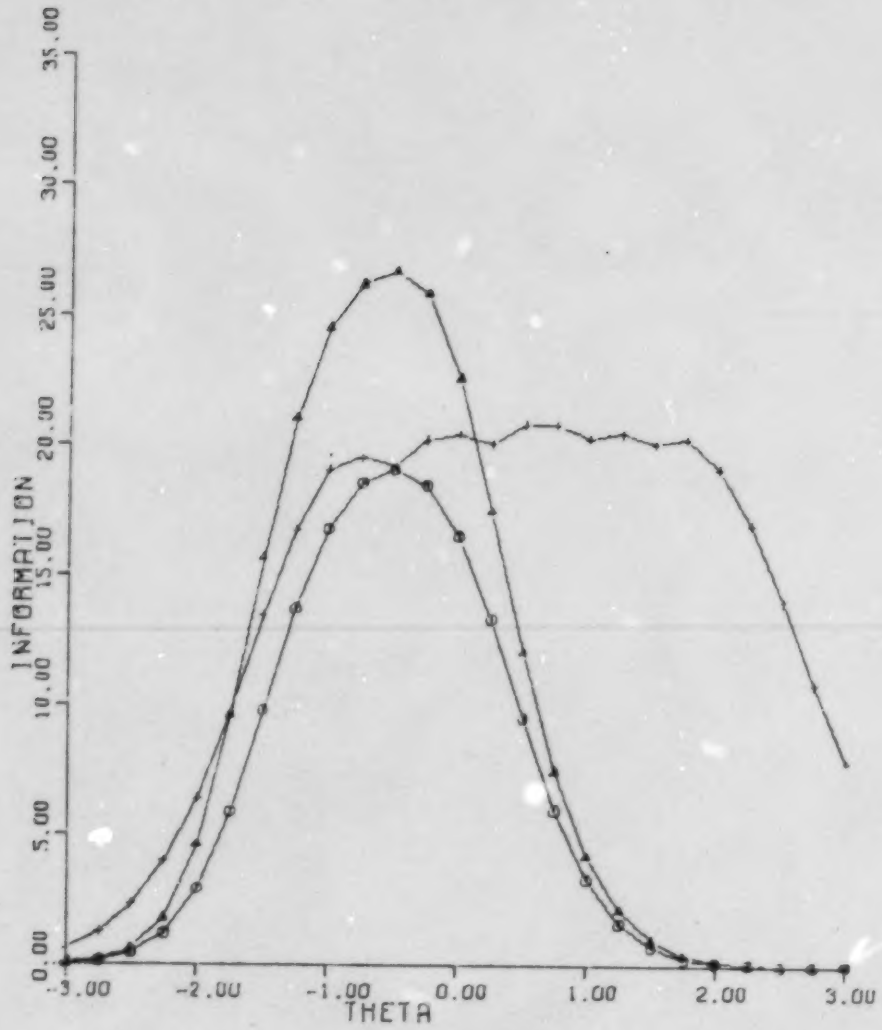

Figure 18. Information functions for perfect tailored and Type III RUC and conventional tests for the item bank defined by Urry's 200 items and guessing pattern G_1 .

over a range of ability values approximately equal to one standard deviation. This seemingly incorrect result is due to the fact that the Type I and II conventional tests contained more items for this range of ability than did the corresponding perfect tailored tests [recall that the length of the conventional tests was set equal to the average (or conditional average) number of items in the tailored tests]. The plots of the information functions for the three varieties of tests for guessing patterns G_2, G_3, G_4 and G_5 were similar enough to those for G_7 that they are not presented in order to save space.

Table 6 contains the comparisons of Types I, II, and III RUC and conventional tests with the perfect tailored tests. All tests summarized in Table 6 were constructed using the five item banks derived from a combination of Koch and Reckase's 72 items and the five guessing patterns. The RUC tests constructed from the Koch and Reckase items were all more informative on the average than the perfect tailored tests with the exception of the Type I and II RUC tests drawn from the item bank defined by G_5. The Type I conventional tests yielded average test information that was about 90 percent of the average information for perfect tailored tests across the five item banks based on the Koch and Reckase items (the actual range was from 88 to 92 percent). The Type II conventional tests produced average test information functions that were 94 to 98 percent of the average test information for the perfect tailored tests. The Type III conventional tests were 95 to 99 percent as efficient as the perfect tailored tests -- as measured by the ratio of average test information functions.

Figures 19, 20 and 21 display the information functions for the perfect tailored and Types I, II and III RUC and conventional tests for the item bank defined by combining the a_g and b_g values for the 72 items compiled by Koch and Reckase with the guessing pattern G_1 (equivalent plots for patterns G_2-G_5 are given in Appendix A).

Table 6

Average Information and Efficiency Ratios
for Perfect Tailored Tests, RUC Tests and Standard Tests.
Item Banks based on Koch & Reckase (1978).

Guessing Pattern	Perfect Tailored Tests (Avg. No. Items)	RUC Tests	Conventional Tests	Efficiency[1] Ratio (RUC Test)	Efficiency Ratio (Conventional Tests)
(1) For full ability distribution (Type I RUC and Conventional Tests)					
G_1	14.3253 (24)	17.5090	12.6779	1.22	0.89
G_2	9.8066 (25)	10.2180	8.6024	1.04	0.88
G_3	16.0274 (22)	19.9012	14.7397	1.24	0.92
G_4	13.5684 (25)	15.4560	12.0105	1.14	0.89
G_5	7.7855 (25)	7.2010	6.8171	0.92	0.88
(2) For above average ability			(Type II RUC and Conventional Tests)		
G_1	13.2451 (25)	15.8599	12.5655	1.20	0.95
G_2	9.2966 (25)	9.8729	8.7087	1.06	0.94
G_3	14.9442 (24)	18.2427	14.5779	1.22	0.98
G_4	12.4538 (25)	14.1316	11.7218	1.13	0.94
G_5	7.5161 (25)	7.3206	7.0280	0.97	0.94
(3) For below average ability			(Type III RUC and Conventional Tests)		
G_1	15.9503 (23)	22.4957	15.3600	1.41	0.96
G_2	10.7525 (25)	12.6672	10.2629	1.18	0.95
G_3	17.4967 (19)	24.4866	17.3205	1.40	0.99
G_4	15.2453 (25)	19.8709	14.6785	1.30	0.96
G_5	8.4139 (25)	8.5867	8.0282	1.02	0.95

[1]Ratio of average information in tests with fixed items (RUC or Conventional) to average information in tailored tests.

The information functions for the RUC tests (Types I, II and III) exceeded the information function for perfect tailored tests over a broader range of ability compared to the equivalent RUC tests based on Urry's item data and pattern G_1. Also the information functions for the conventional tests and perfect tailored tests were more similar for the Koch and Reckase items (see figures 19, 20 & 21) than were the equivalent functions for items based on the Urry data (see Figures 16, 17 & 18). The information functions for the perfect tailored tests constructed using G_1 and the Koch-Reckase item data were highly peaked in comparison to the same functions based on G_1 and Urry's item data. All five sets of perfect tailored tests based on Urry's original item bank were remarkably flat in the range $-2 \leq \theta \leq 2$ -- with the possible exception of the item banks defined by G_2.

The final task in Phase II was a complete repetition of the preceding investigation of Types I, II and III conventional tests vs. perfect tailored tests using the actual (estimated) values of the guessing parameter, c_g, derived by Urry (1974) and Koch & Reckase (1978) for their calibrated item banks. Recall that in the earlier research done using these two item banks, the actual c_g values were ignored, and the first elements of the five guessing patterns were imposed upon the calibrated values of a_g and b_g. This was done because the actual values of the guessing patterns, necessary for evaluating RUC tests, were unknown for the Urry and Koch-Reckase item banks. As was done earlier, the two actual item banks were used to construct two sets of perfect tailored tests (one for each item bank) for each of 49 discrete values of the ability parameter, θ. Types I, II and III conventional tests were drawn from the two item banks; the lengths of these tests were set equal to the average, or conditional average, number of items in the appropriate perfect tailored tests. The average test information for perfect tailored tests was computed

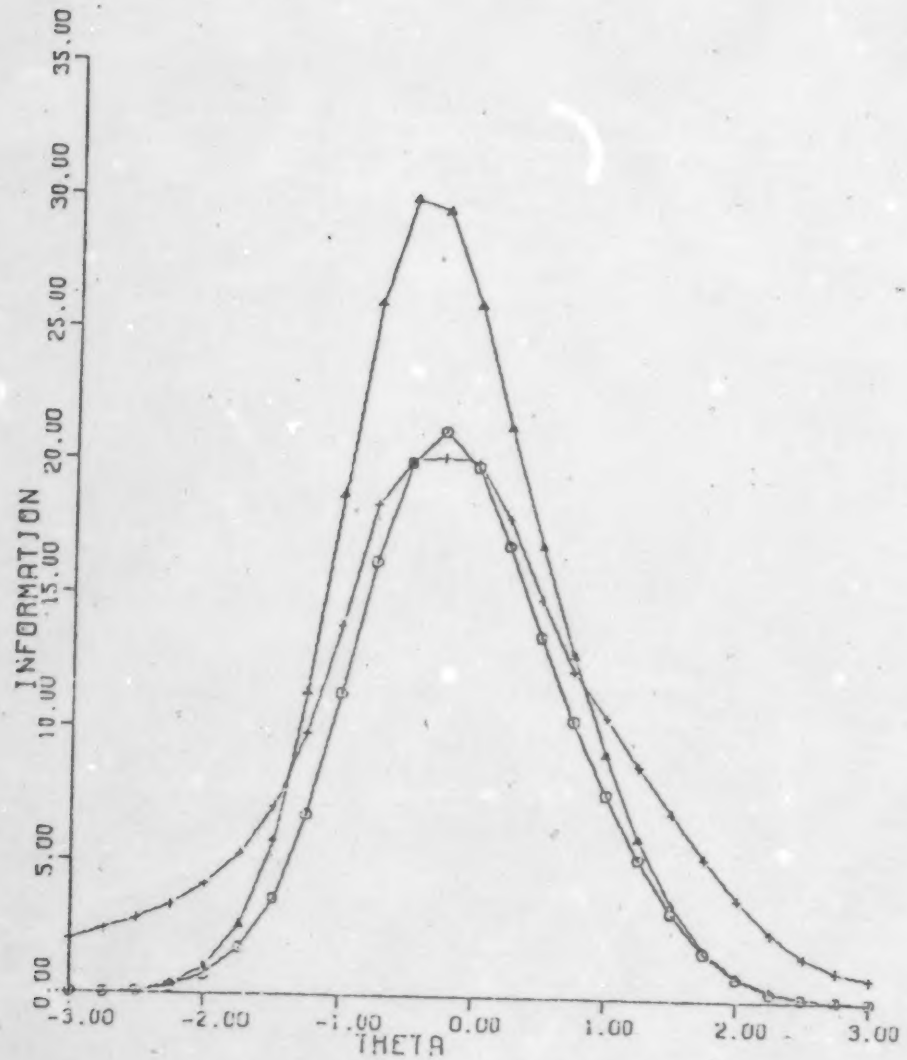

Figure 19. Information functions for perfect tailored and Type I RUC and conventional tests for the item bank defined by Koch & Reckase's 72 items and guessing pattern G_1 .

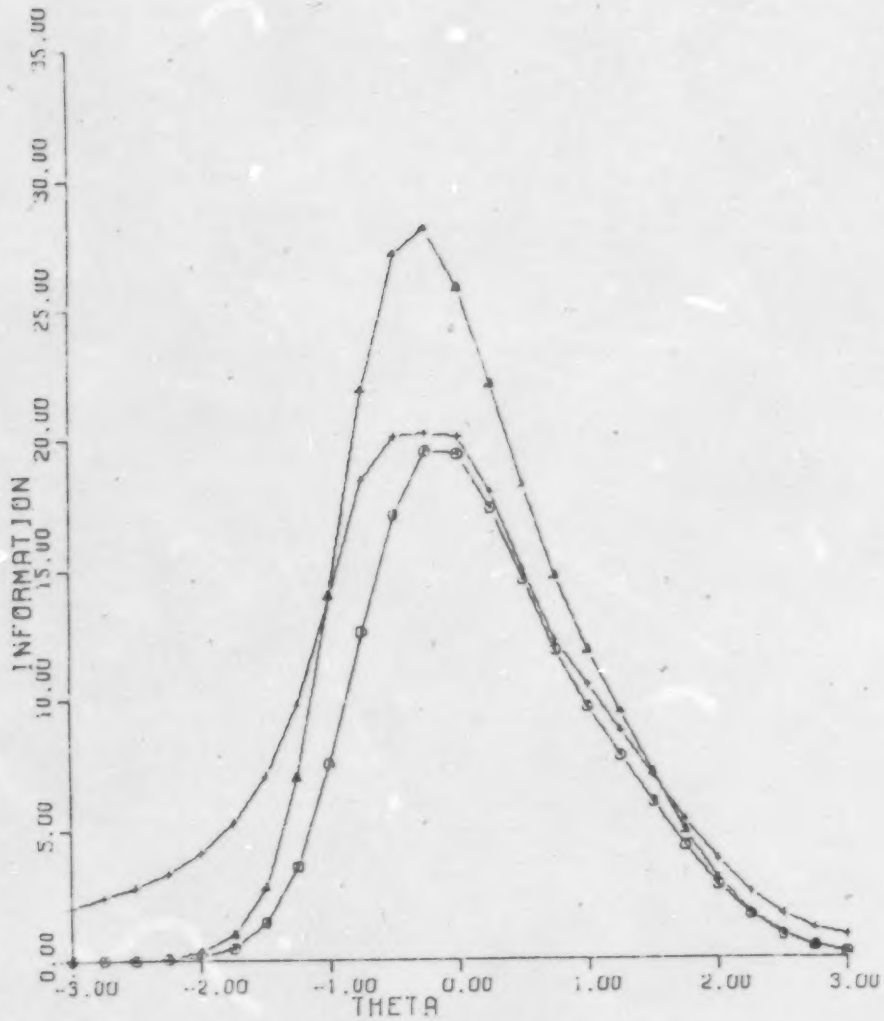

Figure 20. Information functions for perfect tailored and Type II RUC and conventional tests for the item bank defined by Koch and Reckase's 72. items and guessing pattern G_1 .

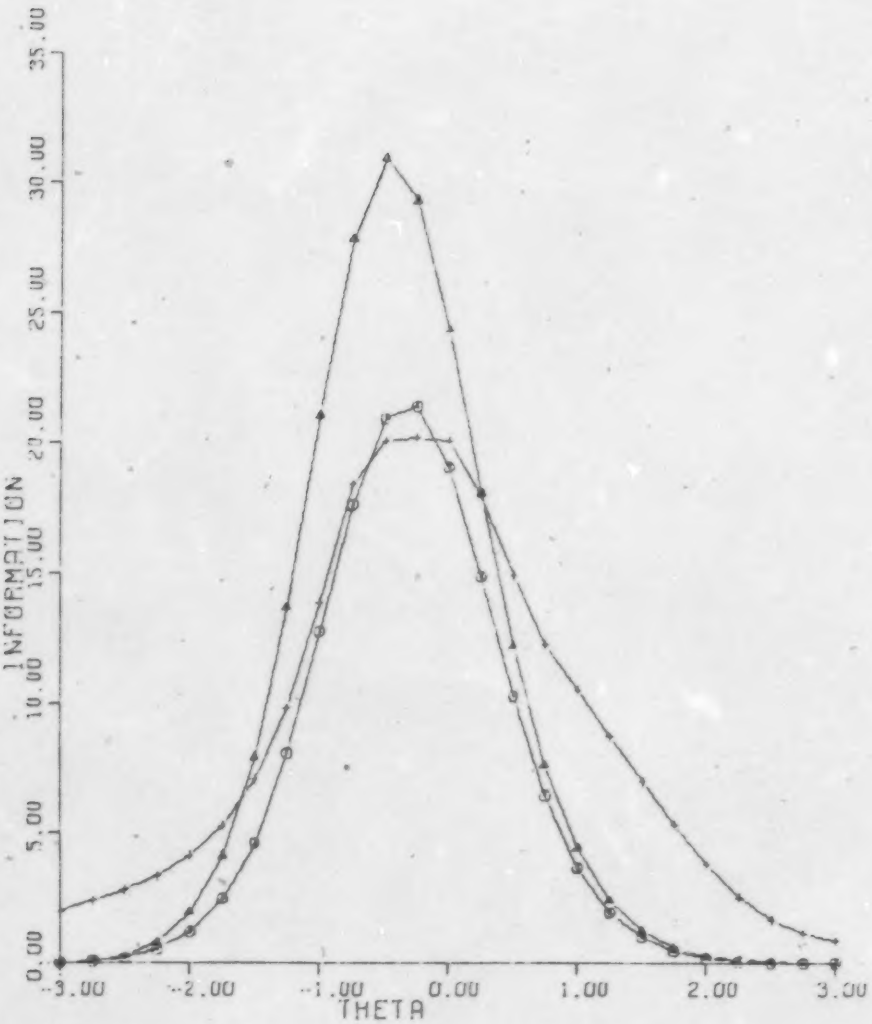

Figure 21. Information functions for perfect tailored and Type III RUC and conventional tests for the item bank defined by Koch & Reckase's 72 items and guessing pattern G_1 .

for the entire ability distribution, above average ability and below average ability for comparison with the average test information for Types I, II and III conventional tests. The Type I conventional tests were determined by choosing the $\bar{n}_I$ items that maximized $\hat{\mu}[I(\theta)]$ [see (21)]. The Type II conventional tests consisted of the $\bar{n}_{II}$ items that maximized $\hat{\mu}[I(\theta|\theta_i \geq 0)]$ [see (22)], and finally the Type III conventional tests consisted of the $\bar{n}_{III}$ items maximizing $\hat{\mu}[I(\theta|\theta_i \leq 0)]$ [see (23)].

Table 7 displays the comparisons of the three types of conventional tests with perfect tailored tests for the two actual item banks. Type I conventional tests compared poorly to the perfect tailored tests from Urry's item bank; however, the Type II and III tests were nearly 90 percent as effective as perfect tailored tests in terms of average test information. The results in Table 7 for Urry's items are extremely close to the equivalent results given in Table 5. The comparisons of conventional with perfect tailored tests in the Koch-Reckase item bank, presented in Table 7, are very similar to the pattern of equivalent outcomes given in Table 6, and indicate that the Types I, II and III conventional tests produced average test information that was virtually identical to that of the perfect tailored tests (especially for Types II and III conventional tests).

The actual test information functions for the Type I, II and III conventional tests and perfect tailored tests derived from Urry's item bank are presented in Figures 22, 23 and 24. In these figures, curves marked with triangles relate to perfect tailored tests, and curves marked with circles relate to conventional tests. In these figures we again see the very flat test information function for the perfect tailored tests constructed from Urry's items -- the information function for the perfect tailored tests is essentially constant for abilities $-2 \leq \theta \leq 2$. In Figures 22 and 23, one will

Table 7

Average Information Functions and Efficiency Ratios for Perfect Tailored and Conventional Tests Constructed from the Item Banks of Urry and Koch & Reckase (Actual c_g values are used)

Urry's Item Bank

Average and Conditional Average Test Information for Perfect Tailored Tests (Average No. of Items)		Average Test Information for Conventional Tests		Efficiency Ratio
Entire Ability Distribution	18.61 ($\bar{n}_I$=22)	Type I	12.75	.69
Above Average Ability	20.08 ($\bar{n}_{II}$=19)	Type II	17.50	.87
Below Average Ability	17.32 ($\bar{n}_{III}$=25)	Type III	15.36	.89

Koch-Reckase Item Bank

Average and Conditional Average Test Information for Perfect Tailored Tests (Average No. of Items)		Average Test Information for Conventional Tests		Efficiency Ratio
Entire Ability Distribution	15.66 ($\bar{n}_I$=22)	Type I	14.07	.90
Above Average Ability	14.41 ($\bar{n}_{II}$=24)	Type II	13.85	.96
Below Average Ability	17.32 ($\bar{n}_{III}$=20)	Type III	17.54	1.01

[1]Ratio of average information in conventional tests to average information perfect tailored tests.

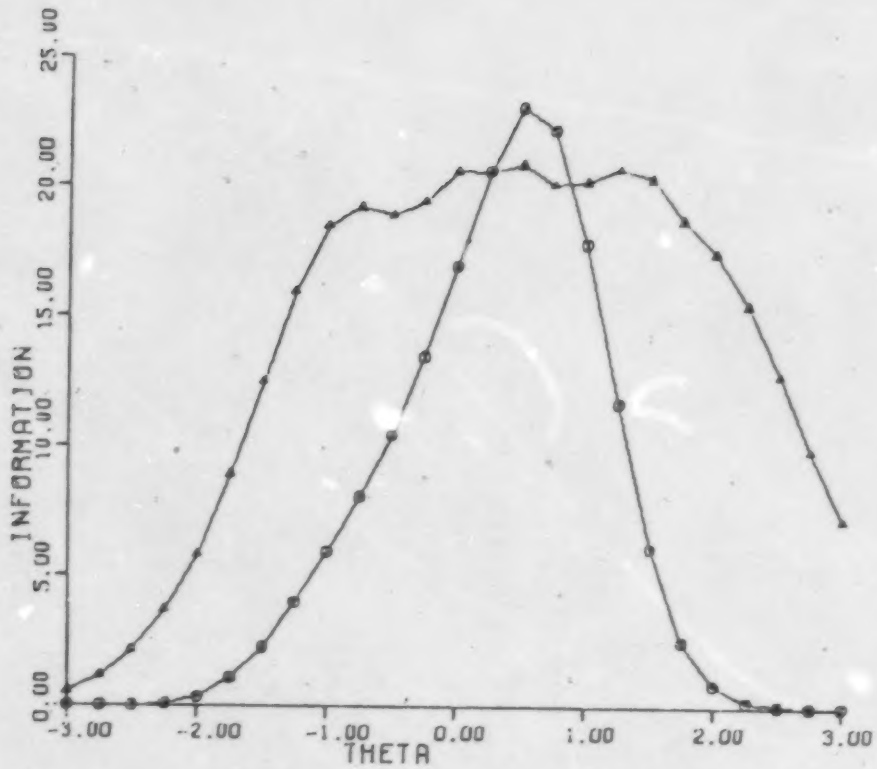

Figure 22. Information functions for perfect tailored and Type I conventional tests. Item bank due to Urry (1974), and actual values of c_g are used.

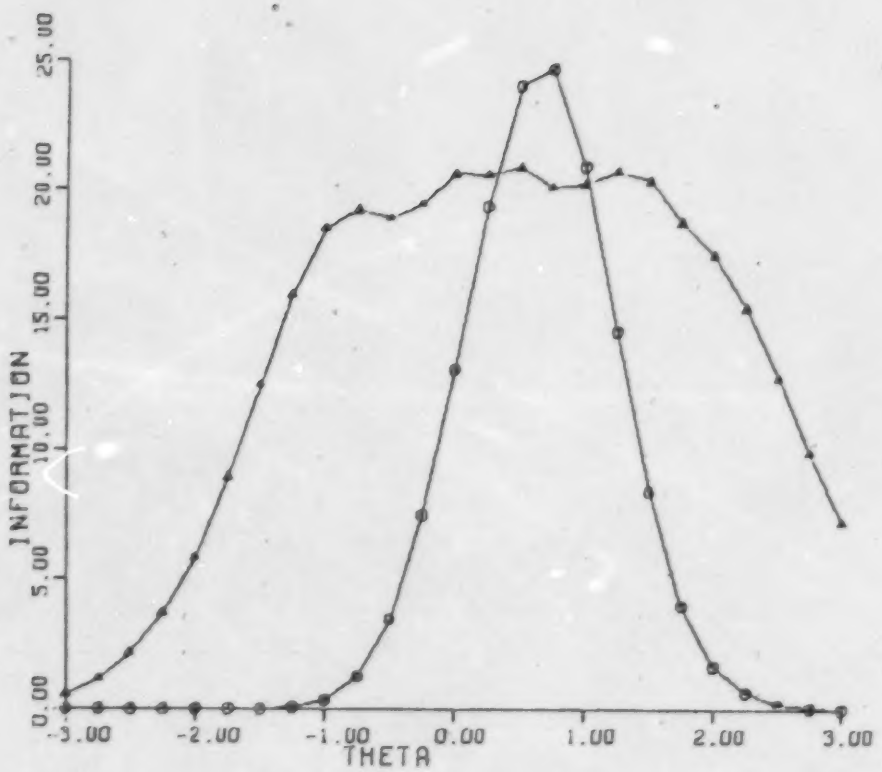

Figure 23. Information functions for perfect tailored and Type II conventional tests. Item bank due to Urry (1974), and actual values of c_g are used.

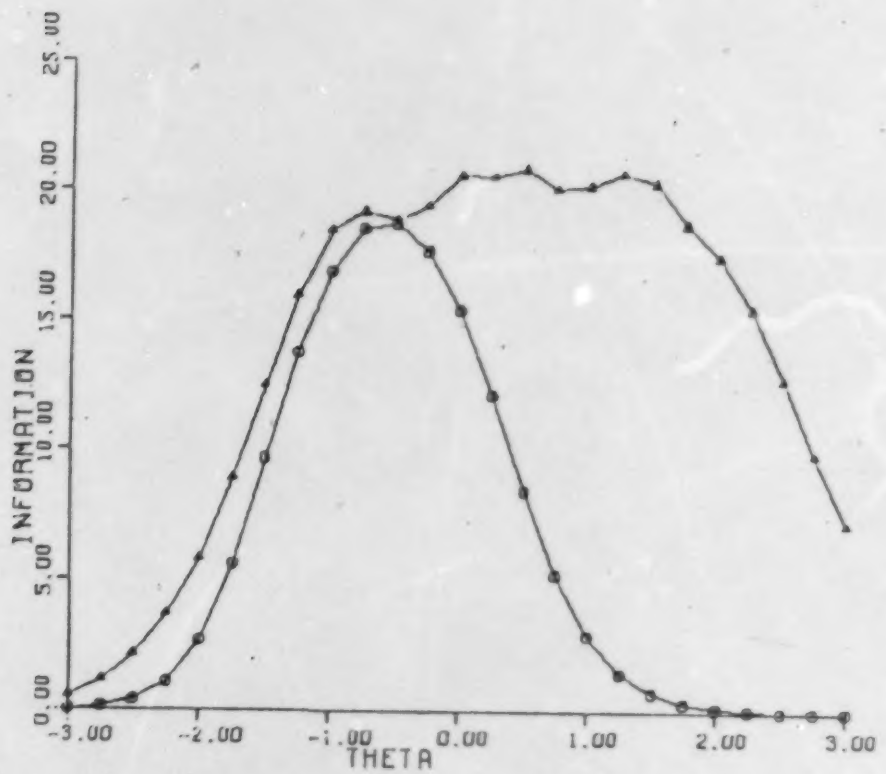

Figure 24. Information functions for perfect tailored and Type III conventional tests. Item bank due to Urry (1974), and actual values of c_g are used.

notice that the test information function for conventional tests exceeds that of the perfect tailored tests for a narrow range of ability. This apparent impossibility stems from the fact that the conventional tests contained more items then the perfect tailored tests for the abilities in this narrow range. Inspection of Figures 23 and 24 may cause one to doubt that the average test information in the conventional tests could possibly be equal to 87 and 89 percent, respectively, of that of the perfect tailored tests. This doubt is alleviated by remembering that the average test information for conventional and tailored tests depicted in Figure 23 is a conditional average based on abilities greater-than or equal-to zero. In the upper-tail of the ability distribution, there is a range of ability (equal to about one-half of a standard deviation) where the conventional test outperformed the perfect tailored tests; this outcome, coupled with the fact the probability densities are vanishingly small in the tail of the ability distribution where the perfect tailored tests substantially outperformed the conventional test, lead to the similarity in (conditional) average information for conventional and tailored tests.

Figures 25, 26 and 27 contain the test information functions for the perfect tailored and conventional tests (Types I, II and III) derived from the Koch-Reckase item bank utilizing the actual values of c_g. Again triangles identify the test information functions for perfect tailored tests and circles identify corresponding functions for conventional tests. For the Koch-Reckase item bank, the test information functions are much more similar to those for the Type I, II and III conventional tests than were the equivalent comparisons for Urry's item bank; this fact is reflected in the similarity of the average test information functions for tailored and conventional tests.

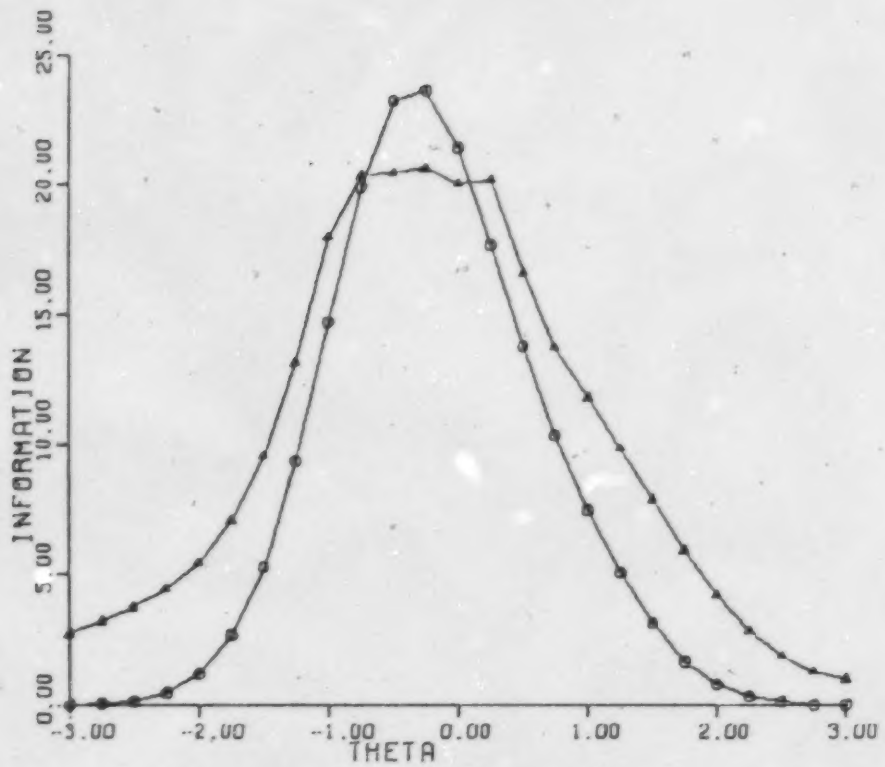

Figure 25. Information functions for perfect tailored and Type I tests. Item bank due to Koch & Reckase (1978), and actual values of c_g are used.

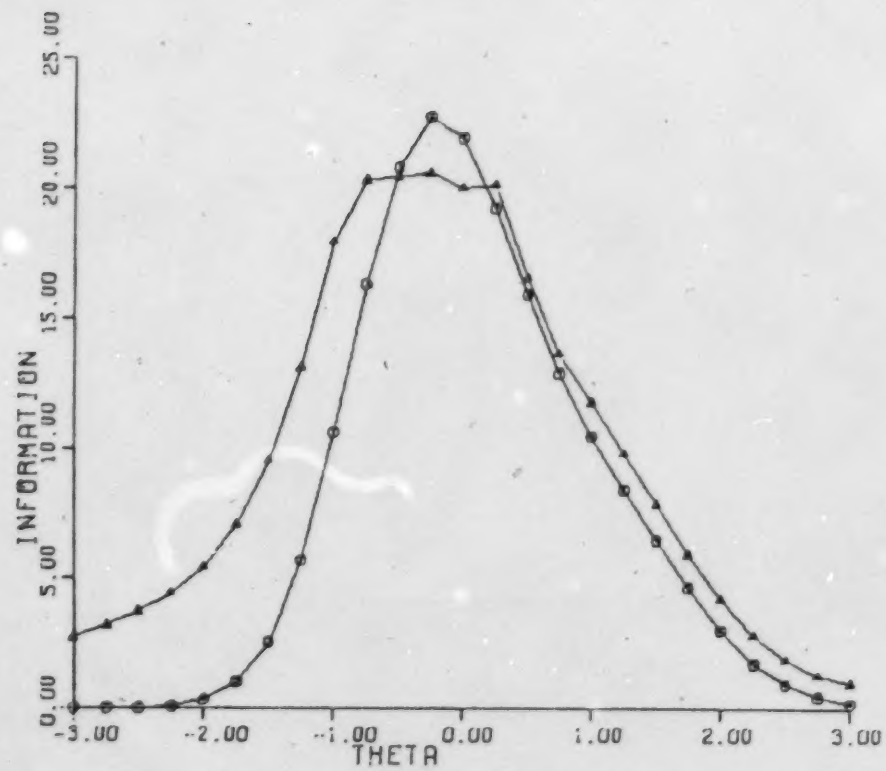

Figure 26. Information functions for perfect tailored and Type II tests. Item bank due to Koch & Reckase (1978), and actual values of c_g are used.

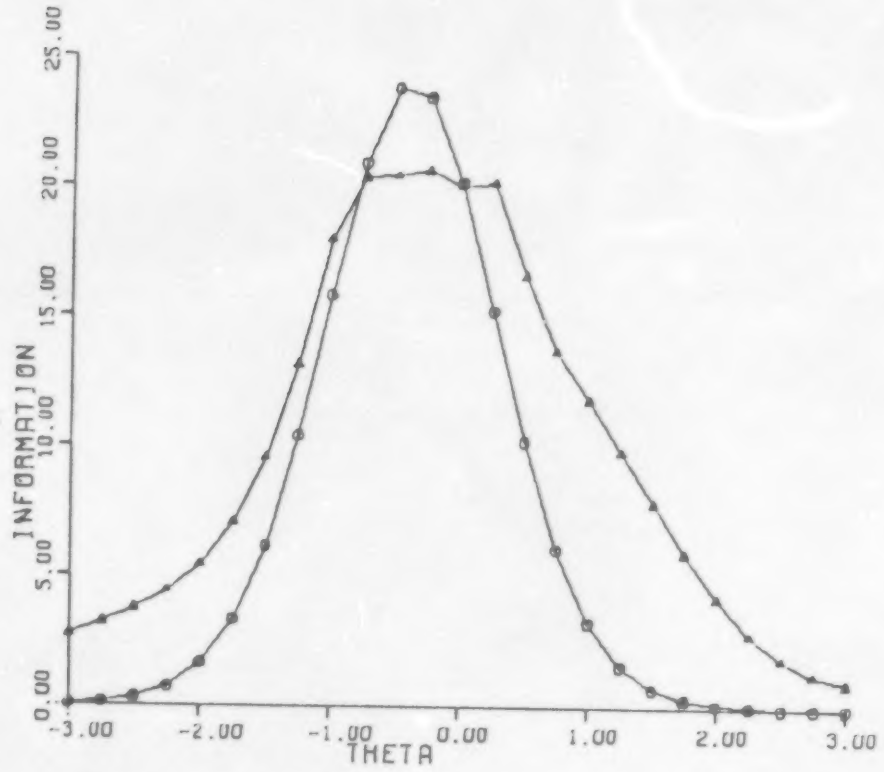

Figure 27. Information functions for perfect tailored and Type III tests. Item bank due to Koch & Reckase (1978), and actual values of c_g are used.

Summary -- Phase II

The item banks of finite length based on collections of items compiled by Urry (1974) and Koch and Reckase (1978) and the five guessing patterns used in Phase I, produced comparative test data for perfect tailored and Types I, II and III RUC and conventional tests that were very similar to those generated by idealized experiments in Phase I. More specifically it was shown that, in certain situations, RUC tests afford greater average test information than perfect tailored tests -- this conclusion assumes, of course, that the model for RUC scores derived by Nicewander, Chang, Doody and Schwarzkopf (1980) is correct, or at least is not a serious distortion of an examinee's behavior when responding repeatedly to multiple-choice items. For the five item banks derived from the 72 Koch-Reckase items and five guessing patterns, the Type II (optimal for above average abilities) and Type III (optimal for below average abilities) conventional tests were at least 94 percent as efficient as perfect tailored tests based on comparisons of the average test information functions. The five item banks generated from Urry's 200 items produced Type II and III conventional tests whose average test information was never less than 83 percent that of the perfect tailored tests, and was as high as 89 percent for two of the Type III tests. A comparison of the average test information for tailored, RUC and conventional tests, using the actual values of c_g from the items calibrated by Urry and Koch and Reckase, produced results very similar to those yielded by substituting the first elements of the five guessing patterns as values of c_g.

Phase III. Estimation of the Loss of Information in Imperfect Tailored Tests Relative to Perfect Tailored Tests.

In Phase III a Monte-Carlo simulation of tailored testing was done using the item banks due to Urry (1974) and Koch and Reckase (1978). The purpose of

this simulation was to compare imperfect tailored tests (in which ability was successively estimated as in an actual tailored testing situation) to the perfect tailored tests constructed when ability is known. In all comparisons of testing methods done in Phases I and II, the tailored tests were perfect (i.e., were constructed to be optimal for known abilities). The purpose of Phase III was to discover the loss of information in actual (imperfect) tailored tests relative to their "perfect" counterparts. In particular, we wanted to answer the question, "Were the RUC and conventional tests used in Phases I and II of this investigation subjected to an unduly harsh evaluation by pitting them against perfect tailored tests?" The Monte Carlo simulation of tailored testing indicated that the answer to the preceding question is "No". The imperfect tailored tests, constructed from the two item banks, yielded nearly identical test information to that of the perfect tailored tests from the same item banks.

A sketch of the Monte Carlo simulation is given below:

1) The "stat-subjects", whose responses were simulated, were assumed to be responding in accord with a three-parameter normal ogive model.

2) Tailored tests for 100 stat-subjects for each of the ability values -2, -1.5, -1, -.5, 0, .5, 1.0, 1.5 and 2 were constructed from the item banks compiled by Urry Koch and Reckase (a total of 1800 stat-subjects were utilized).

3) The first item administered to each stat-subject (from a particular item bank) was one with a b_g value as close to zero as possible and an a_g value as high as possible. If the first item was correct, then an item with a b_g value close to .7 and high a_g was presented; if the first item was incorrect, then the stat-subject was administered an item with $b_g \approx -.7$. If a stat-subject continued with correct responses, the item difficulty (b_g) was increased by steps of approximately .7 until an item was incorrectly answered. Likewise for stat-subjects who persisted with incorrect responses, item difficulty was decreased by a fixed-step size as close to .7 as possible until an item was correctly answered. This process continued until each stat-subject had at least one correct response and at least one incorrect response.

4) Once a stat-subject had compiled a response vector containing at least one correct and one incorrect response, a maximum likelihood estimate of the stat-subject's ability was accomplished via a computer search of the log likelihood function. The next item administered was that which yielded the greatest information at the estimated ability value. This process of maximum likelihood ability estimation and optimal item selection was continued for each stat-subject until: a) the test information function had attained a value of at least 20 or, b) 25 items had been administered.

5) In order to determine whether a stat-subject answered an item correctly or incorrectly, $P_g(\theta) = c_g + (1-c_g)\Phi[L_g(\theta)]$ was computed for the subject-item combination, and a random number drawn from a uniform (0,1) random number generator. If the random number was less-than or equal-to $P_g(\theta)$, the item was recorded as correct; otherwise, the item was recorded as incorrect.

6) In order to summarize the Monte Carlo simulation of the tailored tests for the two actual item banks, the averages of the test information for 100 subjects at each of the nine values of θ were computed and plotted against the test information for the corresponding perfect tailored test.

Figure 28 displays the graphs of the test information functions for the simulated tailored tests and the corresponding perfect tailored tests (from Phase II) for Urry's item bank. (Curves identified by circles are for perfect tailored tests, and curves identified by triangles are for simulated tailored tests.) Notice that the simulated tailored tests, for Urry's bank were nearly as efficient as perfect tailored tests. The loss in information for the simulated tailored tests, relative to perfect tailored tests was never more than two units, and a loss this great occurred only at $\theta = 2$ and $\theta = -1.5$. For $\theta = 1$, the simulated tailored tests had slightly higher information than the corresponding perfect tailored test. The reason for this was that the simulated tests contained more items on the average than the perfect tailored test -- if perfect and fallible tailored tests are equal in length, it must be the case that the information for the perfect tailored test is greater-than or equal-to that of the imperfect tailored test.

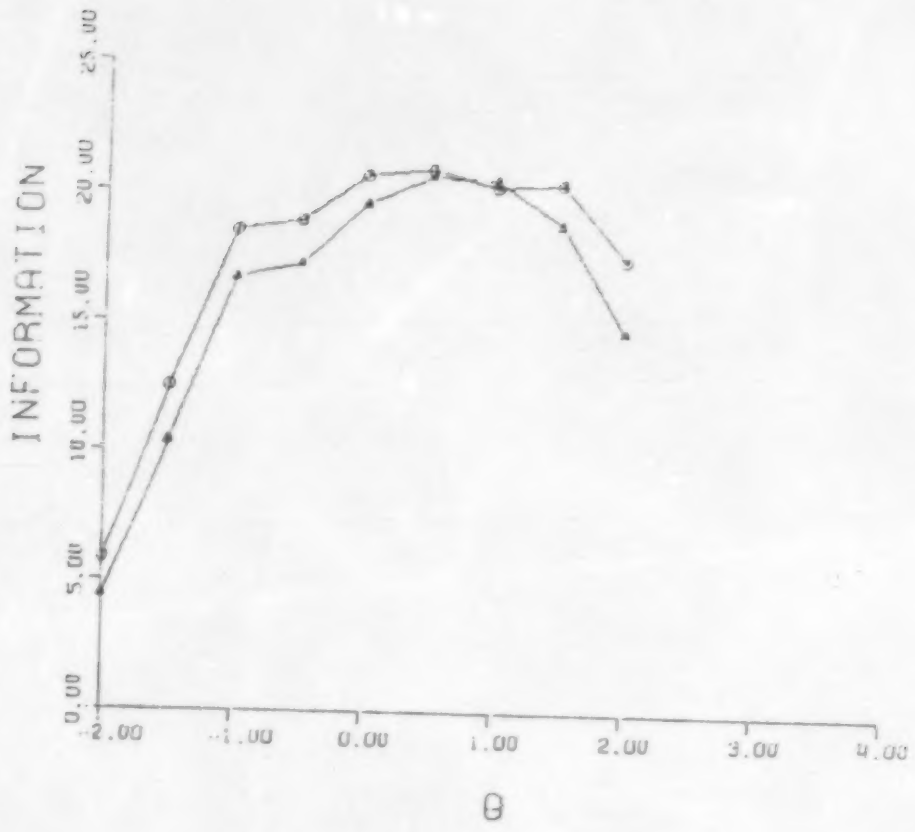

Figure 28. Test information functions for the perfect and simulated tailored tests constructed from Urry's (1974) item bank.

Figure 29 displays the plots of the information functions for the simulated and perfect tailored tests constructed from the Koch-Reckase item bank. As was the case with Urry's item bank, the simulated and perfect tailored tests drawn from this bank yielded very similar information functions indicating that the loss of information due to estimating ability in actual tailored testing is quite small. In Figure 29 it can be seen that the simulated fallible tailored tests never were more than two information units lower than the perfect tailored tests, and a difference as great as two units occurred only at $\theta = -1.5$. At $\theta = 0$, notice that the simulated tailored test had slightly higher information than the perfect tailored test. Again, this was due to the fact that the simulated tailored tests for $\theta = 0$ were slightly longer in length than the corresponding perfect tailored test.

Table 8 lists the number of items in the perfect tailored tests and the average number of items in the simulated fallible tailored tests for the Urry and Koch-Reckase item banks at each of nine abilities. In Table 8 it can be seen that at the ability values for which the simulated tailored tests exceeded the perfect tailored tests in information ($\theta = 1$ for Urry's bank, and $\theta = 0$ for the Koch-Reckase bank) were those ability levels where the simulated tests were approximately two items longer on the average than the equivalent perfect tailored tests.

In order to briefly summarize Phase III of this research, let us state that we were surprised to discover that the imperfect tailored tests, simulated via a Monte Carlo procedure, gave essentially the same test information function as the perfect tailored tests constructed for known abilities. (The simulated tailored tests were slightly longer than their perfect counterparts at certain ability levels in the two item banks utilized for the comparison.) This finding means that the perfect tailored tests, utilized in Phases I and II of this

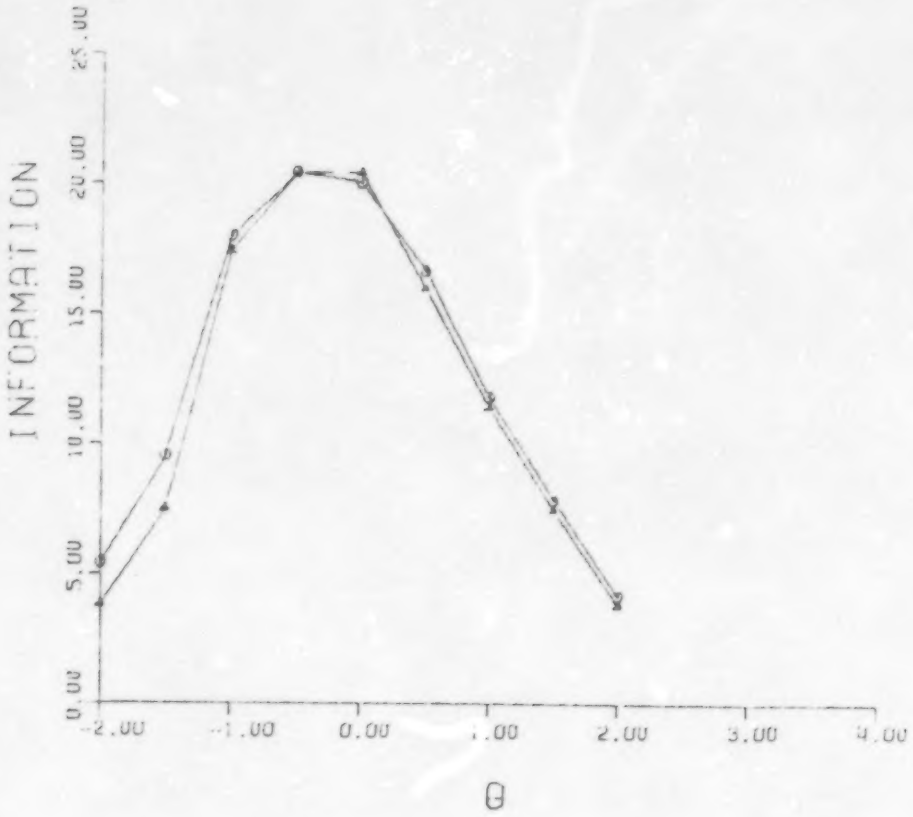

Figure 29. Test information functions for the perfect and simulated tailored tests constructed from Koch & Reckase (1978) item bank.

Table 8

Number of Items in Perfect Tailored Tests
and Average Number of Items in
Simulated Tailored Tests for the Urry
and Koch-Reckase Item Banks at Nine Abilities

	Urry's Item Bank		Koch-Reckase Item Bank	
Ability	Perfect Tailored Tests	Simulated Tailored Tests	Perfect Tailored Tests	Simulated Tailored Tests
-2.0	25	25	25	25
-1.5	25	25	25	25
-1.0	25	25	25	25
-.5	25	25	15	15.97
0	24	24.74	18	19.68
.5	15	16.20	25	25
1.0	15	17.60	25	25
1.5	24	25	25	25
2.0	25	25	25	25

investigation, were not unrealistic criteria against which to judge the measurement characteristics of conventional and RUC tests composed of a fixed set of items for all examinees. It appears that the loss of information due to estimating ability is quite small for actual tailored tests relative to perfect tailored tests. The comparability of perfect tailored tests and simulated tailored tests also suggests that Monte Carlo studies of tailored testing, which assume an underlying three-parameter latent ability model, are probably a waste of computer time. All one really need do in order to discover how a certain item bank will function, when used as an item-pool for tailored testing, is to construct the perfect tailored tests and plot the test information function for a range of abilities. In our studies, constructing the perfect tailored tests from a particular item bank and calculating their test information values took approximately 1/20th the amount of computer time as was required for the Monte Carlo simulation of the fallible tailored tests and their (average) information values.

Summary and Conclusions

A series of analytical, model-based comparisons were made among two, interactive testing methods -- tailored tests and Respond-Until-Correct (RUC) tests composed of a fixed set of items -- and conventional tests (a conventional test was defined as a fixed set of items scored in a binary manner). In the first phase of the inquiry, a model for RUC items was described, and a series of idealized experiments were performed using a three-parameter, normal-ogive model for tailored and conventional tests. The idealized experiments were based on two conceptual entities called "ideal item banks" and "perfect tailored tests", and yielded a series of comparisons among the perfect tailored tests and three types of RUC and conventional tests. In

terms of average test information, it was found that RUC tests were, in many situations, as efficient or more efficient then perfect tailored tests; for a limited number of cases, the conventional tests were competitive with perfect tailored tests in terms of average test information. At the extremes of the ability distribution, all RUC and conventional tests were quite poor providors of information relative to perfect tailored tests.

The second phase of this investigation compared perfect tailored tests and three varieties of RUC and conventional tests using modifications of two actual item banks due to Urry (1974) and Koch and Reckase (1978). Once again it was found that, in terms of average test information, RUC tests were often at least as efficient, or more efficient, than perfect tailored tests. The conventional tests constructed from the finite item banks were nearly as efficient, on the average, as perfect tailored tests in several situations. Even though the RUC and conventional tests, constructed from the finite item banks, often fared well, on the average, relative to perfect tailored tests, they were dismally inefficient at the extremes of the ability distribution compared to perfect tailored tests.

The last phase of the investigation was a Monte-Carlo simulation of "imperfect" tailored testing utilizing the actual item banks compiled by Urry and Koch and Reckase. The purpose of this simulation was to estimate the loss of information, relative to perfect tailored tests, due to the successive estimation of ability in actual tailored testing. It was found that there was essentially no difference between the test information functions for perfect and actual (simulated) tailored tests. Consequently, the perfect tailored tests utilized in the first two phases of this investigation were not unrealistically harsh criteria against which to compare RUC and conventional tests.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Kane, M. and Moloney, J. The effects of guessing on item reliability under answer-until-correct scoring. Applied Psychological Measurement, 1978, 2, No. 1, 41-49.

Koch, W.R. and Reckase, M.D. A live tailored testing study of the one- and three- parameter logistic models. (Research Report 78-1), Columbia: University of Missouri, Tailored Testing Research Laboratory, Department of Educational Psychology, 1978.

Lord, F.M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F.M. and Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Nicewander, W.A., Chang, H.S., Doody, E.N. & Schwarzkopf, A. Some test theory for multiple-choice items whose alternatives are rank-ordered. Unpublished manuscript, Department of Psychology, University of Oklahoma, 1980.

Prater, R. and Nicewander, W.A. Second-guess probabilities for multiple-choice items. Unpublished manuscript, Department of Psychology, University of Oklahoma, 1980.

Pressey, S.L. A simple apparatus which tests and scores-and teaches, School of Society, 1926, 23, 373-376.

Samejima, F. A new family of models for the multiple-choice item. (Research Report 79-4), Knoxville: University of Tennessee, Department of Psychology, 1979.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, No. 2, 233-247.

Urry, V.W. Computer assisted testing: Calibration and evaluation of the verbal ability bank. (Technical Study 74-3), Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, 1974.

Urry, V.W. Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.